

Sri Pratap College

SRINAGAR
LIBRARY

Class No. **R 500.2**
Book No. **S43R**
Accession No. **29551**

Library Sri Pratap College
Srinagar

30 JAN 2006

SRI PRATAP COLLEGE
LIBRARY

Subject Phy (I)

No. 10

DATE LOANED

Acc. No. _____

[illegible]

*Library Sri Pratap College
Srinagar*

**SCIENTIFIC
AMERICAN**

Resource Library

**SELECTIONS READINGS IN THE
Physical Sciences** VOLUME **2**

OFFPRINTS 247-290



W. H. FREEMAN AND COMPANY San Francisco

Library Sri Pratap College
Srinagar

29552

Accession Number.....

Cost Class No....
R
S.D.A. 2
S 43 P

Each of the articles in this volume is available as a separate Offprint. For a complete listing of Offprints available in the Life Sciences, Chemistry, Physics, Technology, Psychology, the Social Sciences, the History and Philosophy of Science, and the Earth Sciences, write to W. H. Freeman and Company, 660 Market Street, San Francisco, California 94104 or to W. H. Freeman and Company, Ltd., Warner House, Folkstone, Kent, England.

Copyright © 1949, 1950, 1951, 1952, 1953, 1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1969, by SCIENTIFIC AMERICAN, INC.

No part of this book may be reproduced by any mechanical, photographic or electronic process, or in the form of a phonographic recording, nor may it be stored in a retrieval system, transmitted, or otherwise copied for public or private use without written permission from the publisher.

Printed in the United States of America.
Library of Congress Catalog Card Number: 78-87738
Standard Book Number: 7167 0991-0

Subject Series:

The *SCIENTIFIC AMERICAN Resource Library* is a multi-volume compilation of more than 700 articles selected from the magazine. These are organized under five subject classifications.

Readings in Earth Sciences (2 volumes)
Readings in Life Sciences (7 volumes)
Readings in Physical Sciences and Technology (3 volumes)
Readings in Psychology (2 volumes)
Readings in Social Sciences (1 volume)

Numbering System:

Each article is numbered and the articles in the volumes are arranged in numerical order. The numbers assigned to each series are:

Earth Sciences, 801-999.
Life Sciences, 1-199 and 1001-1999.
Physical Sciences and Technology, 201-399.
Psychology, 401-599.
Social Sciences, 601-799.

Topic Index:

This index classifies the Readings in Physical Sciences by topic. Note that articles from other subject series that are relevant to the topic are also listed.

Author Index:

The authors of the articles in all five subject series are given. The numbers after the authors' names are the article numbers, not page numbers.

Scientific American Offprints:

Every article in these volumes is published separately in the *SCIENTIFIC AMERICAN* Offprint Series and may be purchased in any quantity. Order by article number and title. The number of each offprint corresponds to the number on each article in the Resource Library. A catalog of *SCIENTIFIC AMERICAN* offprints may be obtained from the publisher.

Additions to the Series:

New titles are added to the *SCIENTIFIC AMERICAN* Offprint Series every month and when enough new articles on a subject are available, a new bound volume will be added to the *SCIENTIFIC AMERICAN Resource Library*.



Contents

VOLUME 1

Topic Index
Author Index

Introduction

Hans A. Bethe	201	What Holds the Nucleus Together?	2
Geoffrey Burbidge and Fred Hoyle	202	Anti-Matter	9
Geoffrey and Margaret Burbidge	203	Stellar Populations	16
Frank B. Cuff, Jr. and L. McD. Schetky	204	Dislocations in Metals	24
Karl K. Darrow	205	The Quantum Theory	30
Harry M. Davis	206	Low Temperature Physics	38
Sergio De Benedetti	207	Mesonic Atoms	49
Freeman J. Dyson	208	Field Theory	55
Albert Einstein	209	On the Generalized Theory of Gravitation	61
William A. Fowler	210	The Origin of the Elements	67
George Gamow	211	The Evolutionary Universe	78
George Gamow	212	The Principle of Uncertainty	86
Murray Gell-Mann and E. P. Rosenbaum	213	Elementary Particles	93
Donald A. Glaser	214	The Bubble Chamber	109
James P. Gordon	215	The Maser	115
Jesse L. Greenstein	216	Dying Stars	125
Robert Hofstadter	217	The Atomic Nucleus	134
Fred Hoyle	218	The Steady-State Universe	145
Donald J. Hughes	219	The Nuclear Reactor as a Research Instrument	149
P. M. Hurley	220	Radioactivity and Time	157
Henry F. Ivey	221	Electroluminescence	163
Abram F. Joffe	222	The Revival of Thermoelectricity	172
Edwin H. Land	223	Experiments in Color Vision	180
Eugene M. Lifshitz	224	Superfluidity	194
Harold Lyons	225	Atomic Clocks	201
Robert E. Marshak	226	Pions	214
B. T. Matthias	227	Superconductivity	220
Maria G. Mayer	228	The Structure of the Nucleus	227
Walter C. Michels	229	The Teaching of Elementary Physics	233
Philip Morrison	230	The Neutrino	242
Philip Morrison	231	The Overthrow of Parity	250
Philip and Emily Morrison	232	The Neutron	260
George E. Pake	233	Magnetic Resonance	271
Wolfgang Panofsky	234	The Linear Accelerator	280
R. E. Peierls	235	Models of the Nucleus	287
Richard F. Post	236	Fusion Power	294
J. S. Prener and D. B. Sullenger	237	Phosphors	306
Edward P. Rosenbaum	238	The Teaching of Elementary Mathematics	312
Bruno Rossi	239	Where Do Cosmic Rays Come From?	322
Allan R. Sandage	240	The Red-Shift	330
Erwin Schrödinger	241	What is Matter?	339
Glenn T. Seaborg and I. Perlman	242	The Synthetic Elements I	346

Glenn T. Seaborg and Albert Ghiorso	243	The Synthetic Elements II	357
Emilio Segrè and Clyde C. Wiegand	244	The Antiproton	369
Frederick Seitz and Eugene P. Wigner	245	The Effects of Radiation on Solids	375
Lyman Spitzer, Jr.	246	The Stellarator	381

VOLUME 2

Topic Index

Author Index

S. B. Treiman	247	The Weak Interactions	392
James A. Van Allen	248	Radiation Belts Around the Earth	403
Gregory H. Wannier	249	The Nature of Solids	413
Gart Westerhout	250	The Radio Galaxy	421
Robert R. Wilson	251	Particle Accelerators	430
Herman Yagoda	252	The Tracks of Nuclear Particles	443
John H. Reynolds	253	The Age of the Elements in the Solar System	452
O. C. Wilson	254	A New Scale of Stellar Distances	459
R. Furth	255	The Limits of Measurement	467
Alfred O. C. Nier	256	The Mass Spectrometer	472
Bryce Crawford, Jr.	257	Chemical Analysis by Infrared	477
George A. W. Boehm	258	Titanium	485
Stephen M. Shelton	259	Zirconium	490
Robert L. Fullman	260	The Growth of Crystals	495
Victor F. Weisskopf and E. P. Rosenbaum	261	A Model of the Nucleus	500
Arthur M. Buswell and Worth H. Rodebush	262	Water	507
Charles M. Herzfeld and Arnold M. Bass	263	Frozen Free Radicals	516
George Gamow	264	The Exclusion Principle	522
B. J. Alder and Thomas E. Wainwright	265	Molecular Motions	531
Boris V. Derjaguin	266	The Force Between Molecules	541
J. D. Bernal	267	The Structure of Liquids	549
Marcus Reiner	268	The Flow of Matter	557
Robert E. Marshak	269	The Nuclear Force	567
Theodore A. Buchhold	270	Applications of Superconductivity	582
Sergio De Benedetti	271	The Mössbauer Effect	592
F. Reif	272	Superfluidity and "Quasi-Particles"	602
George Gamow	273	Gravity	610
Arthur L. Schawlow	274	Optical Masers	621
Sheldon Penman	275	The Muon	632
Roy A. Keller	276	Gas Chromatography	643
H. A. Steinherz and P. A. Redhead	277	Ultrahigh Vacuum	654
D. S. Heesch	278	Radio Galaxies	667
J. E. Kunzler and Morris Tanenbaum	279	Superconducting Magnets	677
John F. Brown, Jr.	280	Inclusion Compounds	686
H. S. Feder and A. E. Spencer	281	Telephone Switching	695
Gerard K. O'Neill	282	The Spark Chamber	707
Philip Morrison	283	Neutrino Astronomy	716
Olexa-Myron Bilaniuk	284	Semiconductor Particle-Detectors	724
D. Nelson Limber	285	The Pleiades	732
Edel Wasserman	286	Chemical Topology	742
Harold Berger	287	Neutron Radiography	750
Robert L. Sproull	288	The Conduction of Heat in Solids	759
Carleen Maley Hutchins	289	The Physics of Violins	767
R. D. Hill	290	Resonance Particles	781

VOLUME 3

Topic Index

Author Index

J. Bronowski	291	The Clock Paradox	794
P. A. M. Dirac	292	The Evolution of the Physicist's Picture of Nature	8
Glenn T. Seaborg and Arnold R. Fritsch	293	The Synthetic Elements III	811
Arthur L. Schawlow	294	Advances in Optical Masers	822
Owen Gingerich	295	The Solar System Beyond Neptune	834
G. F. Chew, Murray Gell-Mann and Arthur H. Rosenfeld	296	Strongly Interacting Particles	842
J. E. Amoore, James W. Johnston Jr., and Martin Rubin	297	The Stereochemical Theory of Odor	862
William H. Pierce	298	Redundancy in Computers	871
Gerald Oster and Yasunori Nishijima	299	Moiré Patterns	878
Emmett N. Leith and Juris Upatnieks	300	Photography by Laser	889
Eugene P. Wigner	301	Violations of Symmetry in Physics	902
Stewart E. Miller	302	Communication by Laser	913
George C. Pimental	303	Chemical Lasers	923
Marshall Gates	304	Analgesic Drugs	933
Geoffrey Burbidge and Fred Hoyle	305	The Problem of the Quasi-Stellar Objects	940
Leo L. Beranek	306	Noise	954
L. K. Runnels	307	Ice	964
Geoffrey Eglinton and Melvin Calvin	308	Chemical Fossils	973
Martin Gardner	309	Can Time Go Backward?	987
Werner H. Wahl and Henry H. Kramer	310	Neutron-Activation Analysis	995
Hannes Alfvén	311	Antimatter and Cosmology	1004
Anatol Rapoport	312	The Use and Misuse of Game Theory	1012
Marvin L. Minsky	313	Artificial Intelligence	1022
Herman F. Mark	314	Giant Molecules	1032
Giulio Natta	315	Precisely Constructed Polymers	1043
Kip S. Thorne	316	Gravitational Collapse	1053
W. Ehrenberg	317	Maxwell's Demon	1063
Bela Julesz	318	Texture and Visual Perception	1072
Jerome B. Wiesner and Herbert F. York	319	National Security and the Nuclear-Test Ban	1083
R. B. Merrifield	320	The Automatic Synthesis of Proteins	1092
R. S. Shankland	321	The Michelson-Morley Experiment	1106
Edward L. Ginzton and William Kirk	322	The Two-Mile Electron Accelerator	1114
Gerard K. O'Neill	323	Particle Storage Rings	1124
Leon M. Lederman	324	The Two-Neutrino Experiment	1133
Sir Lawrence Bragg	325	X-Ray Crystallography	1144
Allen V. Astin	326	Standards of Measurement	1158

Topic Index

(This index includes not only the article in the Physical Sciences but also relevant articles from other subject series.)

ASTROPHYSICS AND COSMOLOGY

- 102. Harrison Brown THE AGE OF THE SOLAR SYSTEM
- 202. Geoffrey Burbidge & Fred Hoyle ANTI-MATTER
- 203. Geoffrey & Margaret Burbidge STELLAR POPULATIONS
- 210. William A. Fowler THE ORIGIN OF THE ELEMENTS
- 211. George Gamow THE EVOLUTIONARY UNIVERSE
- 216. Jesse L. Greenstein DYING STARS
- 218. Fred Hoyle THE STEADY-STATE UNIVERSE
- 239. Bruno Rossi WHERE DO COSMIC RAYS COME FROM?
- 240. Allan R. Sandage THE RED-SHIFT
- 248. James A. Van Allen RADIATION BELTS AROUND THE EARTH
- 250. Gart Westerhout THE RADIO GALAXY
- 253. John H. Reynolds THE AGE OF THE ELEMENTS IN THE SOLAR SYSTEM
- 254. O. C. Wilson A NEW SCALE OF STELLAR DISTANCES
- 278. D. S. Heesch RADIO GALAXIES
- 283. Philip Morrison NEUTRINO ASTRONOMY
- 285. D. Nelson Limber THE PLEIADES
- 295. Owen Gingerich THE SOLAR SYSTEM BEYOND NEPTUNE
- 305. Geoffrey Burbidge & Fred Hoyle THE PROBLEM OF THE QUASI-STELLAR OBJECTS
- 309. Martin Gardner CAN TIME GO BACKWARD?
- 311. Hannes Alfvén ANTIMATTER AND COSMOLOGY
- 316. Kip S. Thorne GRAVITATIONAL COLLAPSE
- 801. Robert S. Dietz ASTROBLEMES
- 802. Virgil E. Barnes TEKTITES
- 851. Robert Jastrow ARTIFICIAL SATELLITES AND THE EARTH'S ATMOSPHERE
- 858. Sir Charles Wright THE ANTARCTIC AND THE UPPER ATMOSPHERE
- 873. Desmond King-Hele THE SHAPE OF THE EARTH

HISTORY

- 205. Karl K. Darrow THE QUANTUM THEORY
- 209. Albert Einstein ON THE GENERALIZED THEORY OF GRAVITATION

- 212. George Gamow THE PRINCIPLE OF UNCERTAINTY
- 229. Walter C. Michels THE TEACHING OF ELEMENTARY PHYSICS
- 231. Philip Morrison THE OVERTHROW OF PARITY
- 238. E. P. Rosenbaum THE TEACHING OF ELEMENTARY MATHEMATICS
- 241. Erwin Schrödinger WHAT IS MATTER?
- 255. R. Furth THE LIMITS OF MEASUREMENT
- 289. Carleen Maley Hutchins THE PHYSICS OF VIOLINS
- 291. J. Bronowski THE CLOCK PARADOX
- 292. P. A. M. Dirac THE EVOLUTION OF THE PHYSICIST'S PICTURE OF NATURE
- 309. Martin Gardner CAN TIME GO BACKWARD?
- 311. Hannes Alfvén ANTIMATTER AND COSMOLOGY
- 317. W. Ehrenberg MAXWELL'S DEMON
- 437. Lewis M. Terman ARE SCIENTISTS DIFFERENT?
- 453. Bernard & Judith Mausner A STUDY OF THE ANTI-SCIENTIFIC ATTITUDE

INSTRUMENTATION AND TECHNIQUE

- 46. George Wald EYE AND CAMERA
- 81. William H. Stein & Stanford Moore CHROMATOGRAPHY
- 82. George W. Gray THE ULTRACENTRIFUGE
- 83. George W. Gray ELECTROPHORESIS
- 84. Curtis A. Williams, Jr. IMMUNOELECTROPHORESIS
- 125. Mary A. B. Brazier THE ANALYSIS OF BRAIN WAVES
- 131. Arthur K. Solomon PUMPS IN THE LIVING CELL
- 1040. Robert S. Ledley & Frank H. Ruddle CHROMOSOME ANALYSIS BY COMPUTER
- 1043. Cyrus Levinthal MOLECULAR MODEL-BUILDING BY COMPUTER
- 206. Harry M. Davis LOW TEMPERATURE PHYSICS
- 214. Donald A. Glaser THE BUBBLE CHAMBER
- 219. Donald J. Hughes THE NUCLEAR REACTOR AS A RESEARCH INSTRUMENT

- 220. P. M. Hurley RADIOACTIVITY AND TIME
- 223. Edwin H. Land EXPERIMENTS IN COLOR VISION
- 229. Walter C. Michels THE TEACHING OF ELEMENTARY PHYSICS
- 233. George E. Pake MAGNETIC RESONANCE
- 234. Wolfgang Panofsky THE LINEAR ACCELERATOR
- 238. E. P. Rosenbaum THE TEACHING OF ELEMENTARY MATHEMATICS
- 251. Robert R. Wilson PARTICLE ACCELERATORS
- 252. Herman Yagoda THE TRACKS OF NUCLEAR PARTICLES
- 255. R. Furth THE LIMITS OF MEASUREMENT
- 256. Alfred O. C. Nier THE MASS SPECTROMETER
- 257. Bryce Crawford, Jr. CHEMICAL ANALYSIS BY INFRARED
- 263. Charles M. Herzfeld & Arnold M. Bass FROZEN FREE RADICALS
- 271. Sergio De Benedetti THE MOSSBAUER EFFECT
- 276. Roy A. Keller GAS CHROMATOGRAPHY
- 277. H. A. Steinherz & P. A. Redhead ULTRAHIGH VACUUM
- 279. J. E. Kunzler & Morris Tanenbaum SUPERCONDUCTING MAGNETS
- 281. H. S. Feder & A. E. Spencer TELEPHONE SWITCHING
- 282. Gerard K. O'Neill THE SPARK CHAMBER
- 284. Olexa-Myron Bilaniuk SEMICONDUCTOR PARTICLE-DETECTORS
- 287. Harold Berger NEUTRON RADIOGRAPHY
- 289. Carleen Maley Hutchins THE PHYSICS OF VIOLINS
- 298. William H. Pierce REDUNDANCY IN COMPUTERS
- 299. Gerald Oster & Yasunori Nishijima MOIRE PATTERNS
- 300. Emmett N. Leith & Juris Upatnieks PHOTOGRAPHY BY LASER
- 302. Stewart E. Miller COMMUNICATION BY LASER
- 303. George C. Pimentel CHEMICAL LASERS
- 310. Werner H. Wahl & Henry H. Kramer NEUTRON-ACTIVATION ANALYSIS
- 312. Anatol Rapoport THE USE AND MISUSE OF GAME THEORY
- 313. Marvin L. Minsky ARTIFICIAL INTELLIGENCE

- 318. Bela Julesz TEXTURE AND VISUAL PERCEPTION
- 322. Edward L. Ginzton & William Kirk THE TWO-MILE ELECTRON ACCELERATOR
- 323. Gerard K. O'Neill PARTICLE STORAGE RINGS
- 325. Sir Lawrence Bragg X-RAY CRYSTALLOGRAPHY
- 492. Peter F. Ostwald ACOUSTIC METHODS IN PSYCHIATRY
- 496. Alphonse Chapanis PSYCHOLOGY AND THE INSTRUMENT PANEL
- 503. Burke M. Smith THE POLYGRAPH
- 510. Oliver G. Selfridge & Ulric Neisser PATTERN RECOGNITION BY MACHINE
- 516. E. Llewellyn Thomas MOVEMENTS OF THE EYE

MATTER IN BULK

- 204. Frank B. Cuff, Jr. & L. McD. Schetky DISLOCATIONS IN METALS
- 224. Eugene M. Lifshitz SUPERFLUIDITY
- 227. B. T. Matthias SUPERCONDUCTIVITY
- 245. Frederick Seitz & Eugene P. Wigner THE EFFECTS OF RADIATION ON SOLIDS
- 249. Gregory H. Wannier THE NATURE OF SOLIDS
- 260. Robert L. Fullman THE GROWTH OF CRYSTALS
- 262. Arthur M. Buswell & Worth H. Rodebush WATER
- 265. B. J. Alder & Thomas E. Wainwright MOLECULAR MOTIONS
- 266. Boris V. Derjaguin THE FORCE BETWEEN MOLECULES
- 267. J. D. Bernal THE STRUCTURE OF LIQUIDS
- 268. Marcus Reiner THE FLOW OF MATTER
- 272. F. Reif SUPERFLUIDITY AND "QUASI-PARTICLES"
- 288. Robert L. Sproull THE CONDUCTION OF HEAT IN SOLIDS
- 307. L. K. Runnels ICE
- 325. Sir Lawrence Bragg X-RAY CRYSTALLOGRAPHY

NUCLEUS AND ATOM

- 201. Hans A. Bethe WHAT HOLDS THE NUCLEUS TOGETHER?
- 207. Sergio De Benedetti MESONIC ATOMS
- 213. Murray Gell-Mann & E. P. Rosenbaum ELEMENTARY PARTICLES
- 217. Robert Hofstadter THE ATOMIC NUCLEUS
- 226. Robert E. Marshak PIONS
- 228. Maria G. Mayer THE STRUCTURE OF THE NUCLEUS
- 230. Philip Morrison THE NEUTRINO
- 232. Philip & Emily Morrison THE NEUTRON
- 235. R. E. Pelerls MODELS OF THE NUCLEUS

- 242. Glenn T. Seaborg & I. Perlman THE SYNTHETIC ELEMENTS I
- 243. GLENN T. Seaborg & Albert Ghiorso THE SYNTHETIC ELEMENTS II
- 244. Emilio Segrè & Clyde C. Wiegand THE ANTI-PROTON
- 247. S. B. Treiman THE WEAK INTERACTIONS
- 261. Victor F. Weisskopf & E. P. Rosenbaum A MODEL OF THE NUCLEUS
- 269. Robert E. Marshak THE NUCLEAR FORCE
- 275. Sheldon Penman THE MUON
- 283. Philip Morrison NEUTRINO ASTRONOMY
- 290. R. D. Hill RESONANCE PARTICLES
- 293. Glenn T. Seaborg & Arnold R. Fritsch THE SYNTHETIC ELEMENTS: III
- 296. Geoffrey F. Chew, Murray Gell-Mann & Arthur H. Rosenfeld STRONGLY INTERACTING PARTICLES
- 301. Eugene P. Wigner VIOLATIONS OF SYMMETRY IN PHYSICS
- 310. Werner H. Wahl & Henry H. Kramer NEUTRON-ACTIVATION ANALYSIS
- 324. Leon M. Lederman THE TWO-NEUTRINO EXPERIMENT

ORGANIC AND BIOCHEMISTRY

- 7. Paul Doty PROTEINS
- 8. Louis F. Fieser STEROIDS
- 10. Joseph S. Fruton PROTEINS
- 15. David E. Green ENZYMES IN TEAMS
- 16. David E. Green THE METABOLISM OF FATS
- 23. K. U. Linderstrom-Lang HOW IS A PROTEIN MADE?
- 28. A. E. Mirsky THE CHEMISTRY OF HEREDITY
- 31. Linus Pauling, Robert B. Corey & Roger Hayward THE STRUCTURE OF PROTEIN MOLECULES
- 35. Francis O. Schmitt GIANT MOLECULES IN CELLS AND TISSUES
- 42. E. O. P. Thompson THE INSULIN MOLECULE
- 54. F. H. C. Crick NUCLEIC ACIDS
- 67. David E. Green THE SYNTHESIS OF FAT
- 68. Mahlon B. Hoagland NUCLEIC ACIDS AND PROTEINS
- 69. Albert L. Lehninger ENERGY TRANSFORMATION IN THE CELL
- 75. Daniel I. Arnon THE ROLE OF LIGHT IN PHOTOSYNTHESIS
- 80. William H. Stein & Stanford Moore THE CHEMICAL STRUCTURE OF PROTEINS
- 85. John D. Roberts ORGANIC CHEMICAL REACTIONS
- 91. Albert L. Lehninger HOW CELLS TRANSFORM ENERGY
- 92. Vincent G. Allfrey & Alfred E. Mirsky HOW CELLS MAKE MOLECULES

- 100. Martin D. Kamen TRACERS
- 101. Philip H. Abelson PALEOBIOCHEMISTRY
- 122. J. A. Bassham THE PATH OF CARBON IN PHOTOSYNTHESIS
- 123. F. H. C. Crick THE GENETIC CODE
- 128. Robert L. Sinsheimer SINGLE-STRANDED DNA
- 153. Marshall W. Nirenberg THE GENETIC CODE: II
- 155. Frederic Verzár THE AGING OF COLLAGEN
- 156. Christian de Duve THE LYSOSOME
- 157. Edward O. Wilson PHEROMONES
- 165. Renato Baserga & Walter E. Kisielewski AUTOBIOGRAPHIES OF CELLS
- 166. Alick Isaacs FOREIGN NUCLEIC ACIDS
- 169. H. O. J. Collier ASPIRIN
- 170. Earl Frieden THE CHEMISTRY OF AMPHIBIAN METAMORPHOSIS
- 171. Alexander Rich POLYRIBOSOMES
- 180. Wolfgang Beermann & Ulrich Clever CHROMOSOME PUFFS
- 183. S. Spiegelman HYBRID NUCLEIC ACIDS
- 186. Sarah Clevenger FLOWER PIGMENTS
- 189. Martin Jacobson & Morton Beroza INSECT ATTRACTANTS
- 190. Maynard A. Amerine WINE
- 193. Heinz Fraenkel-Conrat THE GENETIC CODE OF A VIRUS
- 196. M. F. Perutz THE HEMOGLOBIN MOLECULE
- 198. Hans Neurath PROTEIN-DIGESTING ENZYMES
- 1008. Jean-Pierre Changeux THE CONTROL OF BIOCHEMICAL REACTIONS
- 1012. Emile Zuckerkandl THE EVOLUTION OF HEMOGLOBIN
- 1013. Eric H. Davidson HORMONES AND GENES
- 1016. Eugene I. Rabinowitch & Govindjee THE ROLE OF CHLOROPHYLL IN PHOTOSYNTHESIS
- 1020. Alfred Champagnat PROTEIN FROM PETROLEUM
- 1030. John Cairns THE BACTERIAL CHROMOSOME
- 1033. Robert W. Holley THE NUCLEOTIDE SEQUENCE OF A NUCLEIC ACID
- 1040. Robert S. Ledley & Frank H. Ruddle CHROMOSOME ANALYSIS BY COMPUTER
- 1041. Luigi Gorini ANTIBIOTICS AND THE GENETIC CODE
- 1043. Cyrus Levinthal MOLECULAR MODEL-BUILDING BY COMPUTER
- 1052. F. H. C. Crick THE GENETIC CODE: III
- 1055. David C. Phillips THE THREE-DIMENSIONAL STRUCTURE OF AN ENZYME MOLECULE
- 1057. Aharon Gibor ACETABULARIA: A USEFUL GIANT CELL
- 1061. Philip C. Hanawalt & Robert H. Haynes THE REPAIR OF DNA

1064. Franklin C. McLean BONE
1068. N. W. Pirie ORTHODOX AND UNORTHODOX METHODS OF MEETING WORLD FOOD NEEDS
1074. Charles Yanofsky GENE STRUCTURE AND PROTEIN STRUCTURE
1075. Ruth Hubbard & Allen Kropf MOLECULAR ISOMERS IN VISION
1076. Paul R. Ehrlich & Peter H. Raven BUTTERFLIES AND PLANTS
1077. Bernard W. Agranoff MEMORY AND PROTEIN SYNTHESIS
1078. Carroll M. Williams THIRD-GENERATION PESTICIDES
1081. Elijah Adams BARBITURATES
1082. Trevor Robinson ALKALOIDS
1083. R. R. Porter THE STRUCTURE OF ANTIBODIES
1085. Anthony Allison LYSOSOMES AND DISEASE
1087. Tsutomu Watanabe INFECTIOUS DRUG RESISTANCE
1092. Brian F. C. Clark & Kjeld A. Marcker HOW PROTEINS START
1101. Efraim Racker THE MEMBRANE OF THE MITOCHONDRION
1103. Leonard Hayflick HUMAN CELLS AND AGING
1110. R. D. Preston PLANTS WITHOUT CELLULOSE
1111. Johannes van Overbeek THE CONTROL OF PLANT GROWTH
1122. Sir Solly Zuckerman HORMONES
1124. Arthur Kornberg THE SYNTHESIS OF DNA
280. John F. Brown, Jr. INCLUSION COMPOUNDS
286. Edel Wasserman CHEMICAL TOPOLOGY
297. John E. Amore, James W. Johnston, Jr., & Martin Rubin THE STEREOCHEMICAL THEORY OF ODOR
304. Marshall Gates ANALGESIC DRUGS
308. Geoffrey Eglinton & Melvin Calvin CHEMICAL FOSSILS
314. Herman F. Mark GIANT MOLECULES
315. Giulio Natta PRECISELY CONSTRUCTED POLYMERS
320. R. B. Merrifield THE AUTOMATIC SYNTHESIS OF PROTEINS
325. Sir Lawrence Bragg X-RAY CRYSTALLOGRAPHY
446. Harold E. Himwich THE NEW PSYCHIATRIC DRUGS
483. Frank Barron, Murray E. Jarvik & Sterling Bunnell, Jr. THE HALLUCINOGENIC DRUGS
497. Lawrence Zelic Freedman "TRUTH" DRUGS

PHYSICAL CHEMISTRY: LIQUIDS AND SOLIDS

249. Gregory H. Wannier THE NATURE OF SOLIDS
258. George A. W. Boehm TITANIUM

259. Stephen M. Shelton ZIRCONIUM
260. Robert L. Fullman THE GROWTH OF CRYSTALS
262. Arthur M. Buswell & Worth H. Rodebush WATER
263. Charles M. Herzfeld & Arnold M. Bass FROZEN FREE RADICALS
265. B. J. Adler & Thomas E. Wainwright MOLECULAR MOTIONS
266. Boris V. Derjaguin THE FORCE BETWEEN MOLECULES
267. J. D. Bernal THE STRUCTURE OF LIQUIDS
288. Robert L. Sproull THE CONDUCTION OF HEAT IN SOLIDS
307. L. K. Runnels ICE
854. Edwin Roedder ANCIENT FLUIDS IN CRYSTALS

QUANTUM THEORY AND RELATIVITY

205. Karl K. Darrow THE QUANTUM THEORY
208. Freeman J. Dyson FIELD THEORY
209. Albert Einstein ON THE GENERALIZED THEORY OF GRAVITATION
212. George Gamow THE PRINCIPLE OF UNCERTAINTY
231. Philip Morrison THE OVERTHROW OF PARITY
241. Erwin Schrödinger WHAT IS MATTER?
264. George Gamow THE EXCLUSION PRINCIPLE
273. George Gamow GRAVITY
291. J. Bronowski THE CLOCK PARADOX
292. P. A. M. Dirac THE EVOLUTION OF THE PHYSICIST'S PICTURE OF NATURE
301. Eugene P. Wigner VIOLATIONS OF SYMMETRY IN PHYSICS
309. Martin Gardner CAN TIME GO BACKWARD?
316. Kip S. Thorne GRAVITATIONAL COLLAPSE
321. R. S. Shankland THE MICHELSON-MORLEY EXPERIMENT

TECHNOLOGY

187. Derek H. Fender CONTROL MECHANISMS OF THE EYE
190. Maynard A. Amerine WINE
1020. Alfred Champagnat PROTEIN FROM PETROLEUM
1023. Willem J. Kolff AN ARTIFICIAL HEART INSIDE THE BODY
1036. Joseph B. MacInnis LIVING UNDER THE SEA
1040. Robert S. Ledley & Frank H. Ruddle CHROMOSOME ANALYSIS BY COMPUTER
1043. Cyrus Levinthal MOLECULAR MODEL-BUILDING BY

COMPUTER

215. James P. Gordon THE MASER
221. Henry F. Ivey ELECTROLUMINESCENCE
222. Abram F. Joffe THE REVIVAL OF THERMOELECTRICITY
225. Harold Lyons ATOMIC CLOCKS
236. Richard F. Post FUSION POWER
237. J. S. Prener & D. B. Sullenger PHOSPHORS
246. Lyman Spitzer, Jr. THE STELLARATOR
270. Theodore A. Buchhold APPLICATIONS OF SUPERCONDUCTIVITY
274. Arthur L. Schawlow OPTICAL MASERS
279. J. E. Kunzler & Morris Tanenbaum SUPERCONDUCTING MAGNETS
281. H. S. Feder & A. E. Spencer TELEPHONE SWITCHING
284. Olexa-Myron Bilaniuk SEMICONDUCTOR PARTICLE-DETECTORS
288. Robert L. Sproull THE CONDUCTION OF HEAT IN SOLIDS
294. Arthur L. Schawlow ADVANCES IN OPTICAL MASERS
298. William H. Pierce REDUNDANCY IN COMPUTERS
299. Gerald Oster & Yasunori Nishijima MOIRE PATTERNS
300. Emmett N. Leith & Juris Upatnieks PHOTOGRAPHY BY LASER
302. Stewart E. Miller COMMUNICATION BY LASER
303. George C. Pimentel CHEMICAL LASERS
306. Leo L. Beranek NOISE
310. Werner H. Wahl & Henry H. Kramer NEUTRON-ACTIVATION ANALYSIS
312. Anatol Rapoport THE USE AND MISUSE OF GAME THEORY
313. Marvin L. Minsky ARTIFICIAL INTELLIGENCE
314. Herman F. Mark GIANT MOLECULES
315. Giulio Natta PRECISELY CONSTRUCTED POLYMERS
318. Bela Julesz TEXTURE AND VISUAL PERCEPTION
319. Jerome B. Wiesner & Herbert F. York NATIONAL SECURITY AND THE NUCLEAR-TEST BAN
326. Allen V. Astin STANDARDS OF MEASUREMENT
503. Burke M. Smith THE POLYGRAPH
510. Oliver G. Selfridge & Ulric Neisser PATTERN RECOGNITION BY MACHINE
607. Herbert Butterfield THE SCIENTIFIC REVOLUTION
629. Anne P. Carter THE ECONOMICS OF TECHNOLOGICAL CHANGE
873. Desmond King-Hele THE SHAPE OF THE EARTH

Author Index

(The authors of the articles in all five subject series are given here.)

A Abelson, P. H. 101
Adams, E. 1081
Adams, R. M. 606
Adolph, E. F. 1067
Agranoff, B. W. 1077
Alder, B. J. 265
Alfvén, H. 311
Allen, R. D. 182
Allfrey, V. G. 92
Allison, A. C. 1065, 1085
Amerine, M. A. 190
Amoore, J. E. 297
Anderson, D. L. 855
Andrew, R. J. 627
Arditti, J. 1031
Arnon, D. I. 75
Asch, S. E. 450
Astin, A. V. 326
Axelrod, J. 1015

B Bailey, Jr., H. S. 830
Baker, P. F. 1038
Bales, R. F. 451
Bandura, A. 505
Barnes, V. E. 802
Barnett, S. A. 1060
Barron, F. 432, 483
Bartholomew, G. A. 1120
Bascom, W. 828, 845
Baserga, R. 165
Bass, A. M. 263
Bassett, C. A. L. 1021
Bassham, J. A. 122
Batra, L. R. 1086
Batra, S. W. T. 1086
Beadle, G. W. 1
Bearn, A. G. 150
Beermann, W. 180
Benzer, S. 120
Benzinger, T. H. 129
Beranek, L. L. 306
Berelson, B. 621
Berger, H. 287
Berkowitz, L. 481
Berlyne, D. E. 500
Bernal, J. D. 267
Bernstein, J. 829
Beroza, M. 189
Best, J. B. 149
Bethe, H. A. 201
Bettelheim, B. 433, 439
Biale, J. B. 118
Biddulph, S. & O. 53
Bilaniuk, O. 284
Billingham, R. E. 148
Bitterman, M. E. 490
Blackler, A. W. 94
Blough, D. S. 458
Boehm, G. A. W. 258
Bogert, C. M. 1119
Bonner, J. T. 164, 1051
Botelho, S. Y. 194
Bower, T. G. R. 502
Boycott, B. B. 1006
Brachet, J. 90

Brady, J. V. 425
Bragg, L. 325
Braidwood, R. J. 605
Braude, A. I. 177
Braun, A. C. 1024
Brazier, M. A. B. 125
Brett, J. R. 1019
Brindley, G. S. 1089
Broadbent, D. E. 467
Bronowski, J. 291
Broom, R. 832
Brown, H. 102
Brown, Jr., J. F. 280
Brues, C. T. 838
Buchhold, T. A. 270
Bullen, K. E. 804
Bunnell, Jr., S. 483
Burbidge, G. 202, 203, 305
Burbidge, M. 203
Burnet, M. 2, 3, 78, 138
Bustad, L. K. 1045
Buswell, A. M. 262
Butler, R. A. 426
Butler, W. L. 107
Butterfield, H. 607

C Cairns, J. 1030
Calhoun, J. B. 506
Calvin, M. 308
Caplan, N. S. 638
Carey, N. 4
Carr, A. 1010
Carter, A. P. 629
Cattell, R. 475
Ceraso, J. 509
Champagnat, A. 1020
Changeaux, J. 1008
Chapanis, A. 496
Chapman, C. B. 1011
Chew, G. F. 296
Clark, B. F. C. 1092
Clark, J. D. 820
Clarke, C. A. 1126
Clements, J. A. 142
Clevenger, S. 186
Clever, U. 180
Cobb, W. C. 109
Cockrill, R. 1088
Cohen, J. 427, 489
Colbert, E. H. 806
Cole, L. C. 144
Collier, H. O. J. 132, 169
Comer, J. P. 633
Comroe, Jr., J. H. 1034
Constantinides, P. F. 4
Cooper, C. F. 1099
Coopersmith, S. 511
Corey, R. B. 31
Crary, A. P. 857
Crawford, Jr., B. 257
Crick, F. H. C. 5, 54, 123, 1052
Crombie, A. C. 184
Crow, J. F. 55
Csapo, A. 163
Cuff, Jr., F. B. 204

D Darrow, K. K. 205
Davidson, E. H. 1013
Davis, H. M. 206
Dawkins, M. J. R. 1018
De Benedetti, S. 207, 271
de Duve, C. 156
Deering, R. A. 143
Deevey Jr., E. S. 608, 811, 834, 844
de Heinzelin, J. 613
Delbrück, M. 1104
Delbrück, M. B. 1104
Denenberg, V. H. 478
Derjaguin, B. V. 266
DeVore, I. 614
Dietz, R. S. 801, 866
Dilger, W. C. 1049
Dirac, P. A. M. 292
Dobzhansky, T. 6, 609
Doty, P. 7
Doumani, G. A. 863
Dowling, J. E. 1053
Downs, R. J. 107
Dulbecco, R. 1069
Dyson, F. J. 208

E Eaton, J. W. 440
Ebert, J. D. 56
Eccles, J. C. 65, 1001
Echlin, P. 1044, 1105
Edgar, R. S. 1004, 1079
Edwards, R. G. 1047
Eglinton, G. 308
Ehrenberg, W. 317
Ehrlich, P. R. 1076
Eibl-Eibesfeldt, I. 470
Einstein, A. 209
Eiseley, L. C. 108, 846
Ellison, W. D. 817
Elsasser, W. M. 825
Emerson, R. 115
Emiliani, C. 815
Emlen, J. T. 1054
Epstein, R. H. 1004
Ericson, D. B. 856
Esch, H. 1071
Etkin, W. 1042
Eysenck, H. J. 477

F Fairbridge, R. W. 805
Fantz, R. L. 459
Feder, H. S. 281
Fender, D. H. 187
Ferster, C. B. 484
Festinger, L. 472
Field, W. O. 809
Fieser, L. F. 8
Fischberg, M. 94
Fisher, A. E. 485
Fisher, R. L. 814
Flyger, V. 1102
Fowler, W. A. 210
Fraenkel-Conrat, H. 9, 193
Franzini-Armstrong, C. 1007
Fraser, D. 59
Freedman, L. Z. 497

- Freedman, R. 621
Freimer, E. H. 1028
French, J. D. 66
Frieden, E. 170
Fritsch, A. R. 293
Fromm, E. 495
Fruton, J. S. 10
Fuhrman, F. A. 1080
Fullman, R. L. 260
Funkenstein, D. H. 428
Furth, J. J. 119
Furth, R. 255
- G** Gamow, G. 211, 212, 264, 273
Gardner, M. 309
Gates, D. M. 1029
Gates, M. 304
Gazzaniga, M. S. 508
Gell-Mann, M. 213, 296
Gerard, R. W. 11
German III, J. L. 150
Gesell, A. 401
Ghiorso, A. 243
Gibor, A. 1057
Gibson, E. J. 402
Gilbert, P. W. 127
Gilliard, E. T. 1098
Gingerich, O. 295
Ginzton, E. L. 322
Glaessner, M. F. 837
Glaser, D. A. 214
Glass, H. B. 1062
Gleitman, H. 479
Goldstein, K. 445
Gordon, J. P. 215
Gorini, L. 1041
Govindjee 1016
Grant, V. 12
Gray, G. W. 13, 14, 82, 83, 103, 1063
Gray, J. 1113
Green, D. E. 15, 16, 67
Greenberg, B. 1017
Greenstein, J. L. 216
Gregory, R. L. 517
Griffin, D. R. 1121
Gross, J. 88
Gruenberg, E. M. 441
Guhl, A. M. 471
Gurdon, J. B. 1128
Guttman, N. 403
- H** Haagen-Smit, A. J. 404, 618
Hadorn, E. 1127
Hall, C. S. 434
Hammond, E. C. 126
Hanawalt, P. C. 1061
Harlow, H. F. 415, 429, 473
Harlow, M. F. 415, 473
Harris, C. S. 507
Hartline, H. K. 99
Hasler, A. D. 411
Hayashi, T. 97
Hayflick, L. 1103
Haynes, R. H. 1061
Hayward, R. 31
Heeschen, D. S. 278
Heezen, B. C. 807
Heirtzler, J. R. 875
Heiskanen, W. A. 812
Held, R. 494
Heron, W. 430
Herzfeld, C. M. 263
Hess, E. H. 416, 493
Hildebrand, M. 1114
Hill, R. D. 290
Himwich, H. E. 446
- Hoagland, M. B. 68
Hockett, C. F. 603
Hodgson, E. S. 1048
Hoffenberg, M. 611
Hofstadter, R. 217
Hokin, L. E. & M. R. 1022
Hollaender, A. 57
Holley, R. W. 1033
Holter, H. 96
Hong, S. K. 1072
Horne, R. W. 147
Horowitz, N. H. 17
Hotchkiss, R. D. 18
Howells, W. W. 604, 630
Hoyle, F. 202, 218, 305
Hubbard, R. 1075
Hubel, D. H. 168
Hudson, J. W. 1120
Hughes, D. J. 219
Hull, D. 1018
Hurley, P. M. 220, 874
Hurwitz, J. 119
Hutchins, C. M. 289
Huxley, H. E. 19, 1026
Huxley, J. 616
Hydén, H. 134
Hyman, H. H. 623
- I** Ingram, V. M. 104
Irving, L. 1032
Irwin, O. C. 417
Isaacs, A. 87, 166
Ittelson, W. H. 405
Ivey, H. F. 221
- J** Jackson, D. D. 442, 447, 468
Jacob, F. 50, 89
Jacobs, W. P. 116
Jacobson, L. F. 514
Jacobson, M. 189
Janowitz, M. 433
Janssen, R. E. 826
Jarvik, M. E. 483
Jastrow, R. 851
Jensen, D. 1035
Joffe, A. F. 222
Johansen, K. 1125
Johnson, C. G. 173
Johnston, Jr., J. W. 297
Jones, J. C. 1106
Jouvet, M. 504
Julesz, B. 318
- K** Kalish, H. I. 403
Kalmus, H. 406
Kamen, M. D. 100
Katona, G. 452
Katz, B. 20, 98
Kaufman, L. 462
Kay, M. 816
Kellenberger, E. 1058
Keller, R. A. 276
Kellogg, C. E. 821
Kendrew, J. C. 121
Kennedy, D. 162, 1073
Kettlewell, H. B. D. 842
Keynes, R. D. 58
Kilpatrick, F. P. 405
King-Hele, D. 873
Kirk, W. 322
Kisielewski, W. E. 165
Kleitman, N. 431, 460
Knight, C. A. 59
Kohler, I. 465
Kolers, P. A. 487, 512
Kolff, W. J. 1023
- Koller, D. 117
Konigsberg, I. R. 191
Kornberg, A. 1124
Kort, V. G. 860
Kortlandt, A. 463
Kramer, H. H. 310
Krogh, A. 21
Krogman, W. M. 632
Kropf, A. 1075
Kuenen, P. H. 803
Kunzler, J. E. 279
Kylstra, J. A. 1123
- L** Lack, D. 22
Lamb, I. M. 111
Land, E. H. 223
Landsberg, H. E. 824
Langbein, W. B. 869
Langer, W. L. 619
Larsen, J. A. 411
Lederman, L. M. 324
Ledley, R. S. 1040
Lehninger, A. L. 69, 91
Lehrman, D. S. 488
Leighton, A. H. 634
Leith, E. N. 300
Leontief, W. W. 610, 611, 617, 624
Leopold, L. B. 869
Levey, R. H. 188
Levine, J. 435
Levine, S. 436, 498
Levinthal, C. 1043
Lewis, O. 631
Li, C. H. 160
Liddell, H. S. 418
Lifshitz, E. M. 224
Limbaugh, C. 135
Limber, D. N. 285
Linderstrom-Lang, K. U. 23
Lissmann, H. W. 152
Livingston, W. K. 407
Llano, G. A. 865
Loewenstein, W. R. 70
Long, W. E. 863
Lorenz, K. Z. 412
Luria, S. E. 24
Lyons, H. 225
- M** MacInnis, J. B. 1036
MacNeish, R. S. 625
MacNichol, Jr., E. F. 197
Maio, J. J. 1094
Malefijt, A. de W. 639
Malkus, J. S. 847
Mangelsdorf, P. C. 25, 26
Mangin, W. 635
Marcker, K. A. 1092
Mark, H. F. 314
Marshak, R. E. 226, 269
Masserman, J. H. 443
Matthes, G. H. 836
Matthias, B. T. 227
Matyniak, K. A. 515
Mausner, B. & J. 453
Mayer, M. G. 228
Mayerson, H. S. 158
Mazia, D. 27, 93
McCarty, M. 1028
McDermott, W. 612
McDonald, J. E. 839
McElroy, W. D. 141
McLean, F. C. 1064
McNeil, M. 870
McVay, S. 1046
Mellaart, J. 620
Melnick, J. L. 1037

- Melzack, R. 457, 469
Merrifield, R. B. 320
Michels, W. C. 229
Miller, G. A. 419
Miller, S. E. 302
Miller, W. H. 99
Milot, J. 831
Minsky, M. L. 313
Mirsky, A. E. 28, 92, 1109
Mitchell, J. H. 1011
Mittwoch, U. 161
Montagna, W. 1003
Moore, S. 80, 81
Morowitz, H. J. 1005
Morrison, E. 232
Morrison, P. 230, 231, 232, 283
Moscona, A. A. 95
Mrosovsky, N. 513
Muller, H. J. 29
Mulvaney, D. J. 628
Munk, W. 813
Muntz, W. R. A. 179
Murphy, R. C. 864
Myers, J. N. 876
Mykutowycz, R. 1108
- N** Napier, J. 140, 1070
Natta, G. 315
Naylor, A. W. 113
Neisser, U. 486, 510
Neurath, H. 198
Newell, N. D. 867
Nichols, J. R. 491
Niederhauser, J. S. 109
Nier, A. O. C. 256
Nirenberg, M. W. 153
Nishijima, Y. 299
Nossal, G. J. V. 199
- O** Olds, J. 30
Oliver, J. 827
O'Neill, G. K. 282, 323
Öpik, E. J. 835
Opler, M. K. 444
Oster, G. 299
Ostwald, P. F. 492
- P** Paige, J. M. 638
Pake, G. E. 233
Panofsky, W. 234
Parducci, A. 518
Pauling, L. 31
Peierls, R. E. 235
Penman, S. 275
Penney, R. L. 1054
Penrose, L. S. 74
Pequegnat, W. E. 853
Perlman, I. 242
Perutz, M. F. 196
Peterson, L. R. 499
Petrunkévitch, A. 1097
Phillips, D. C. 1055
Piaget, J. 420
Piel, G. 413
Pierce, W. H. 298
Pimentel, G. C. 303
Pirie, N. W. 1068
Plass, G. N. 823
Pollard, E. C. 32
Porter, K. R. 1007
Porter, R. R. 1083
Post, R. F. 236
Powers, C. F. 1056
Prenner, J. S. 237
Preston, R. D. 1110
Pritchard, R. M. 466
- Puck, T. T. 33, 71
- R** Rabinowitch, E. I. 34, 1016
Racker, E. 1101
Rahn, H. 1072
Rapoport, A. 312
Rapp, F. 1037
Rasmussen, H. 86
Ratliff, F. 99
Raven, P. H. 1076
Redhead, P. A. 277
Reif, F. 272
Reiner, M. 268
Rensch, B. 421
Revelle, R. 814
Reynolds, J. H. 253
Rich, A. 171
Riesen, A. H. 408
Roberts, J. D. 85
Roberts, W. O. 849
Robertson, A. 1056
Robertson, J. D. 151
Robin, G. de Q. 861
Robinson, T. 1082
Rock, I. 422, 462, 507
Rodebush, W. H. 262
Roedder, E. 854
Roeder, K. D. 1009
Rogers, C. R. 448
Rosenbaum, E. P. 213, 238, 261
Rosenfeld, A. H. 296
Rosenthal, R. 514
Rosenzweig, M. R. 501
Rossi, B. 239
Rothschild, M. 1027
Rubin, H. 185
Rubin, M. 297
Rubin, M. J. 859
Ruddle, F. H. 1040
Runcorn, S. K. 871
Runnels, L. K. 307
Rushton, W. A. H. 139
Ruud, J. T. 1025
- S** Sacher, E. J. 480
Sager, R. 1002
Sahlins, M. D. 602
Salisbury, F. B. 110, 112
Sandage, A. R. 240
Satir, P. 79
Sauer, E. G. F. 133
Savory, T. H. 137, 1039, 1112
Sayre, A. N. 818
Schawlow, A. L. 274, 294
Scheerer, M. 476
Schetky, L. McD. 204
Schmidt-Nielsen, B. 1050
Schmidt-Nielsen, K. 1050, 1096, 1118
Schmitt, F. O. 35
Schneirla, T. C. 413
Scholander, P. F. 172, 1117
Schrödinger, E. 241
Scott, B. I. H. 136
Seaborg, G. T. 242, 243, 293
Segré, E. 244
Seilacher, A. 872
Seitz, F. 245
Selfridge, O. G. 510
Seliger, H. H. 141
Shankland, R. S. 321
Shaw, E. 124
Sheatsley, P. B. 623
Shelton, S. M. 259
Sherif, M. 454
Siekevitz, P. 36
Silvers, W. K. 148
- Simons, E. L. 622, 636
Simpson, D. 640
Singer, M. 105
Sinsheimer, R. L. 128
Skinner, B. F. 423, 461
Slavson, S. R. 449
Smith, B. M. 503
Smith, D. S. 1014
Smith, H. W. 37
Smith, N. G. 1084
Snider, R. S. 38
Sokal, R. R. 1059
Solomon, A. K. 76, 131
Sonneborn, T. M. 39
Southern, H. N. 1095
Speirs, R. S. 176
Spencer, A. E. 281
Sperry, R. W. 72, 174, 1090
Spiegelman, S. 183
Spitzer, Jr., L. 246
Sproull, R. L. 288
Stapleton, G. E. 57
Starr, V. P. 841
Stein, W. H. 80, 81
Steinherz, H. A. 277
Stent, G. S. 40
Stetson, H. C. 808
Stettner, L. J. 515
Steward, F. C. 167
Stewart, S. E. 77
Stommel, H. 810
Storer, J. H. 1115
Stouffer, S. A. 455
Stumpf, P. K. 41
Sullenger, D. B. 237
- T** Taeuber, K. E. 626
Tanenbaum, M. 279
Tanner, J. M. 1091
Taussig, H. B. 1100
Taylor, J. H. 60
Tepper, M. 848
Terman, L. M. 437
Thomas, E. L. 516
Thompson, E. O. P. 42
Thompson, W. R. 469
Thorne, K. S. 316
Thorpe, W. H. 145
Tinbergen, N. 414, 456
Tourtellotte, M. E. 1005
Townsend, M. R. 1102
Treiman, S. B. 247
Turnbull, C. M. 615
Tuttle, O. F. 819
Tyler, A. 43
- U** Underwood, B. J. 482
Upatnieks, J. 300
Urey, H. C. 833
- V** Van Allen, J. A. 248
Van der Kloot, W. G. 146
van Overbeek, J. 1111
Verzár, F. 155
von Békésy, G. 44
von Frisch, K. 130
von Holst, E. 464
von Saint Paul, U. 464
- W** Waddington, C. H. 45
Wahl, W. H. 310
Wainwright, T. E. 265
Wald, G. 46, 47, 48, 61
Walk, R. D. 402
Wallach, H. 409, 474
Walter, W. G. 73

Wannier, G. H. 249
Warden, C. J. 424
Washburn, S. L. 601, 614
Waskow, A. I. 637
Wasserman, E. 286
Watanabe, T. 1087
Wecker, S. C. 195
Weckler, J. E. 844
Weeks, J. R. 178
Weil, R. J. 440
Weiss, E. 18
Weisskopf, V. F. 261
Welty, C. 1116
Wenner, A. M. 181
Went, F. W. 114, 852
Westerhout, G. 250
Wexler, H. 843
Wiegand, C. C. 244

Wiesner, J. B. 319
Wiggers, C. J. 62
Wigglesworth, V. B. 63
Wigner, E. P. 245, 301
Williams, Jr., C. A. 84
Williams, C. M. 49, 1078
Williams, H. 822
Wilson, D. M. 1107
Wilson, E. O. 157
Wilson, J. T. 868
Wilson, O. C. 254
Wilson, R. R. 251
Witkin, H. A. 410
Wittreich, W. J. 438
Wollin, G. 856
Wollman, E. L. 50, 89
Wood, J. E. 1093
Wood, W. B. 1079

Wood, Jr., W. B. 51
Woodcock, A. H. 850
Woodwell, G. M. 159, 1066
Woollard, G. P. 862
Wright, C. 858
Wurtman, R. J. 1015
Wynne-Edwards, V. C. 192

Y Yagoda, H. 252
Yanofsky, C. 1074
York, H. F. 319

Z Zamecnik, P. C. 52
Zimmermann, M. H. 154
Zinder, N. D. 106
Zuckerlandl, E. 1012
Zuckerman, S. 1122
Zweifach, B. W. 64

DATE LOANED

Acc. No. _____

[illegible]

**SCIENTIFIC
AMERICAN**

Resource Library

Physical Sciences 2

THE WEAK INTERACTIONS

by S. B. Treiman

They are now recognized as reflecting a fourth force of nature. The other three are gravity, electromagnetism and the "strong" force which holds together the particles of the atomic nucleus.

Just as physics was engrossed with the force of gravity throughout the 17th and 18th centuries, and with electromagnetic forces in the 19th century, so it was to be expected that physicists in this century would remain preoccupied with nuclear forces. The immensely powerful forces that bind together the elementary particles in the nucleus of the atom are as yet imperfectly understood. Their nature is expressed not only in the enormous energy released in the fusion and fission of atomic nuclei but in the creation of new and mysterious particles in violent nuclear collisions. Nonetheless the nuclear forces have already begun to share the center of the stage—perhaps only temporarily—with an entirely new class of forces. These are the unimaginably weak forces associated with the spontaneous decay and transformation of most of the new particles that so confound the structure of physical theory. Though our acquaintance with these forces is short, it has already led to some of the most disturbing and hopeful developments in modern physics.

We cannot, of course, observe forces. What we observe are interactions—the interaction of relatively substantial bodies in the case of gravity, the interaction of charged bodies and fields in the case of electromagnetic forces and the interaction of subatomic particles in the case of the nuclear and weak forces. Physicists began seriously to conjure with the weak interactions as such only a dozen years ago, when the list of "elementary" particles had already begun to assume alarming length. At present the list of experimentally discovered or theoretically inferred particles stands at 30 [see chart, pages 394 and 395]. None of the late-comers to the list has any part in the constitution of matter in

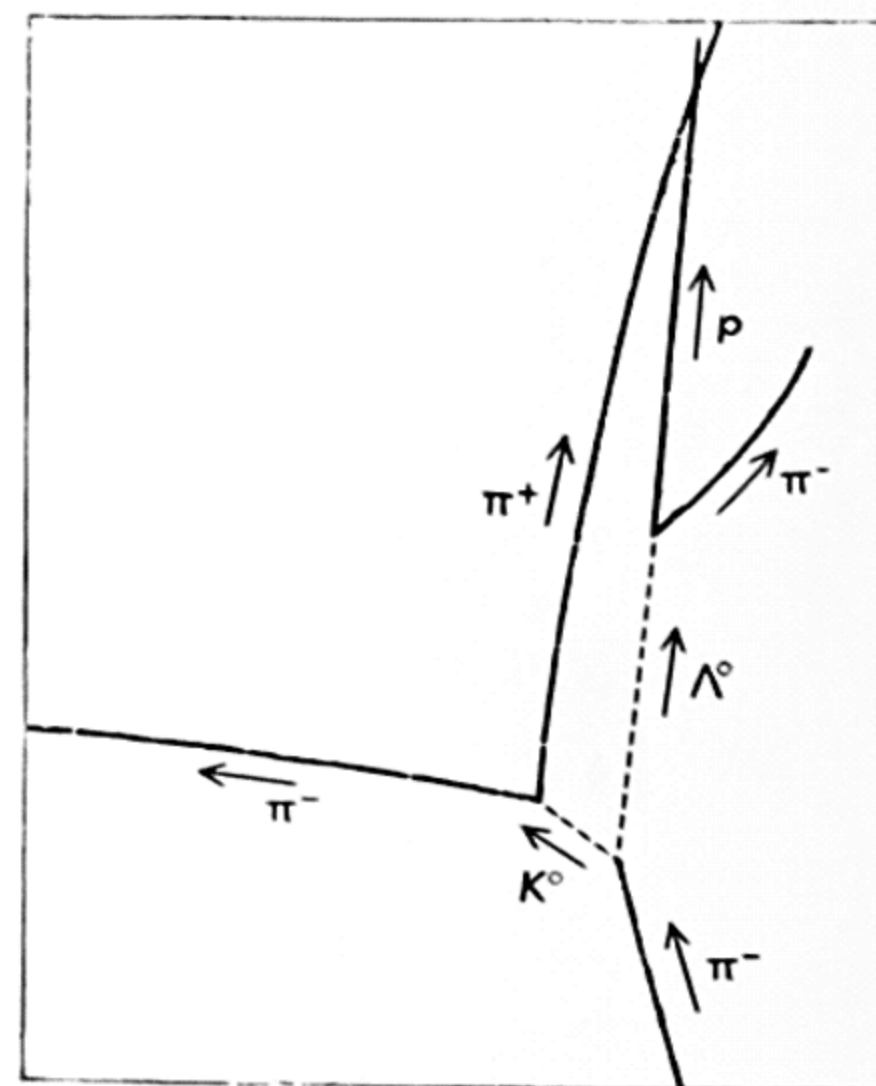
the ordinary sense. All of them are unstable; most of them decay by weak-interaction processes. It was, in fact, the recognition of the weak interactions as the hallmark of one distinct class of particle transformations that gave the first hint of pattern and order in the multiplicity of particles.

More recently certain anomalies observed in the weak interactions inspired the now famous collaboration of Tsung Dao Lee and Chen Ning Yang and of Chien Shiung Wu and her co-workers at the National Bureau of Standards. Their brilliant theoretical and experimental work showed that the weak interactions violate the parity principle. For the first time one of the sacred conservation laws of physics was found inoperable in nature. We can get an idea of the significance of this finding if we try to imagine physics without the law of conservation of energy, or try to think of order without symmetry. A second and equally sacred symmetry principle was toppled at the same time: the symmetry between matter and antimatter. We may yet find a way to resolve nature's apparent disregard for these central propositions of human understanding. If we succeed, it is clear that we will find that order in nature rests upon symmetries subtler than our theories have hitherto imagined.

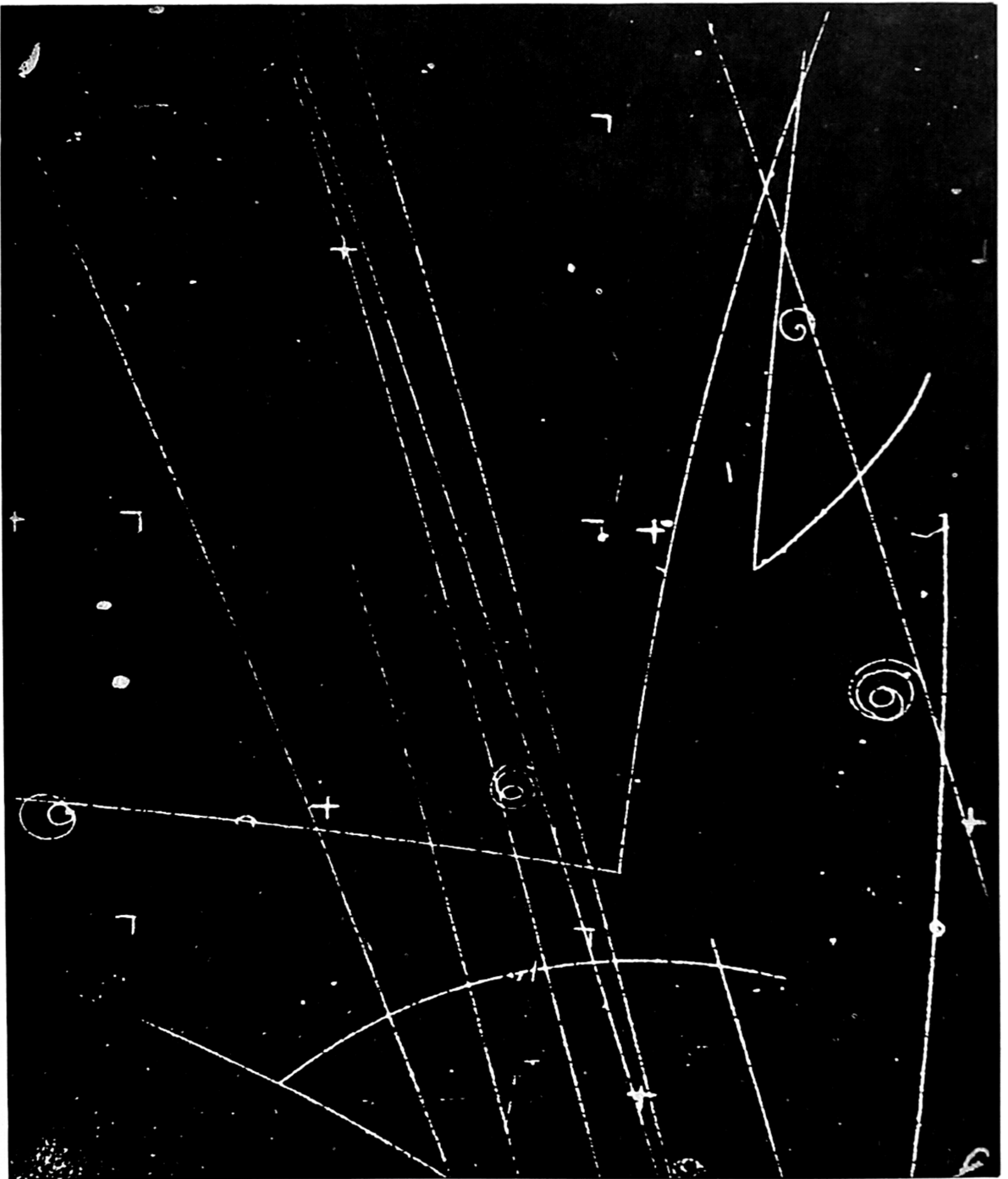
Reaction Rate and Strength

The notion of the weakness or strength of an interaction has, it must be admitted, somewhat medieval overtones. In the broadest sense what the physicist measures is the rate of a reaction: the absolute rate and the rate with respect to other reactions. He measures the rate as a function of the several variables in the process—the

energies of the particles, their momenta and so on. One over-all determinant of rate is the amount of energy available in the reaction. But another and more fundamental determinant that turns up in the equations we invent to describe particle transformations is the intrinsic strength of the reaction. To be sure, these equations are exceedingly complicated, and they can be solved only when we make approximations of dubious validity. What is more, the whole point of our awe at the richness of particle physics is that we are sure our theories are incomplete and inadequate. Nonetheless the characterization of interactions in terms of intrinsic strength is meaningful enough and has uncovered



DECAY of two fundamental particles by a weak-interaction process is illustrated in the bubble-chamber photograph at right, made by Luis W. Alvarez and his colleagues at the



University of California. The events in the photograph are traced in the drawing at left. A high-energy negative pi meson (π^-), produced by the Berkeley Bevatron, enters the chamber at lower right. It strikes a proton in the liquid hydrogen of the bubble chamber,

giving rise to a neutral K meson (K^0) and a lambda particle (Λ^0). Being uncharged, these two particles leave no track. The neutral K meson decays into a negative pi meson and a positive pi meson; the lambda particle, into a proton (p) and a negative pi meson.

patterns and clues that are leading us to deeper insights.

The rates observed in the strong and weak interactions are so widely different as to segregate these interactions unequivocally into two classes of particle reactions. Consider first the strong interactions which predominate in high-energy collisions. In one well-established "strong" process, the collision of a proton (a stable constituent of the atomic nucleus) and a pi meson (an unstable particle that contributes to the powerful forces within the nucleus) yields a lambda particle and a K meson (new unstable particles whose function in nature is still obscure). The time-scale of this reaction can be estimated in the following way. From a variety of experiments we know that the strong forces fall off sharply in space and reach across interparticle distances of no more than 10^{-13} centimeter (.0000000000001 cm.). By observation we know that the particles in the collision are traveling at velocities close to that of light, that is, nearly 3×10^{10} centimeters per second (30,000,000,000 cm./sec.). To find the time interval during which the two particles are close enough to experience their mutual forces we divide the range of the forces by the velocity of the particles. The order of magnitude is 10^{-23} second, or approximately the time it takes light to travel across the diameter of a particle. For the particles to interact in such a brief instant, the forces between them must be immensely strong.

Now let us consider by contrast the time-scale of a weak interaction. We observe that the lambda particle created in the high-energy collision decays into its daughter particles (a proton and a pi meson) in a mean time of about $3 \times$

FUNDAMENTAL PARTICLES that are presently known are listed in the table at right. Negatively charged particles are indicated by a minus; positively charged particles, by a plus; neutral particles, by a zero. The particles in parentheses in the second column are antiparticles. In some cases the antiparticle is indicated by a bar over its symbol; in others, by a sign of charge opposite that of the "particle." The unit of mass in the third column is the mass of the electron. The particles to the right of an arrow in the fourth column are the particles into which the particle to the left of the arrow decays. Where there are alternative modes of decay, they are listed below the arrow. The symbol $<$ indicates "less than"; the symbol \sim , "approximately."

PARTICLE	SYMBOL	MASS	PRINCIPAL MODES OF DECAY	LIFETIME (SECONDS)
PHOTON	γ	0	STABLE	
NEUTRINO	$\nu (\bar{\nu})$	0	STABLE	
ELECTRON	$e^- (e^+)$	1	STABLE	
MU MESON	$\mu^- (\mu^+)$	206	$\mu^- \rightarrow e^- + \nu + \bar{\nu}$	2.22×10^{-6}
PI MESONS	$\pi^- (\pi^+)$	273	$\pi^- \rightarrow \mu^- + \bar{\nu}$	2.56×10^{-8}
			$\pi^- \rightarrow \gamma + \gamma$	
K MESONS	$K^- (K^+)$	264	π^0	$< 10^{-15}$
		967	$\mu^- + \bar{\nu}$	
			$\pi^- + \pi^0$	
			$\pi^- + \pi^+ + \pi^-$	
			$\pi^- + \pi^+ + \pi^0 + \pi^-$	

	K_1^0	~973	$\begin{array}{c} \textcircled{K_1^0} \longrightarrow \textcircled{\pi^+} + \textcircled{\pi^-} \\ \textcircled{\pi^0} + \textcircled{\pi^0} \end{array}$	$\begin{array}{c} \textcircled{\mu^-} + \textcircled{\bar{\nu}} + \textcircled{\pi^0} \\ \textcircled{\pi^+} + \textcircled{\pi^-} \\ \textcircled{\pi^0} + \textcircled{\pi^0} \end{array}$	10^{-10}
	K_2^0	~973	$\textcircled{K_2^0} \longrightarrow \begin{array}{l} \textcircled{\pi^+} + \textcircled{\pi^-} + \textcircled{\pi^0} \\ \textcircled{\pi^0} + \textcircled{\pi^0} + \textcircled{\pi^0} \\ \textcircled{e^-} + \textcircled{\bar{\nu}} + \textcircled{\pi^+} \\ \textcircled{e^+} + \textcircled{\nu} + \textcircled{\pi^-} \\ \textcircled{\mu^-} + \textcircled{\bar{\nu}} + \textcircled{\pi^+} \\ \textcircled{\mu^+} + \textcircled{\nu} + \textcircled{\pi^-} \end{array}$	$\begin{array}{c} \textcircled{\pi^+} + \textcircled{\pi^-} + \textcircled{\pi^0} \\ \textcircled{\pi^0} + \textcircled{\pi^0} + \textcircled{\pi^0} \\ \textcircled{e^-} + \textcircled{\bar{\nu}} + \textcircled{\pi^+} \\ \textcircled{e^+} + \textcircled{\nu} + \textcircled{\pi^-} \\ \textcircled{\mu^-} + \textcircled{\bar{\nu}} + \textcircled{\pi^+} \\ \textcircled{\mu^+} + \textcircled{\nu} + \textcircled{\pi^-} \end{array}$	$\sim 8 \times 10^{-8}$
PROTON	$p(\bar{p})$	1836	STABLE		
NEUTRON	$n(\bar{n})$	1839	$\textcircled{n} \longrightarrow \textcircled{e^-} + \textcircled{p} + \textcircled{\bar{\nu}}$	$\textcircled{e^-} + \textcircled{p} + \textcircled{\bar{\nu}}$	10^{10}
LAMBDA PARTICLE	$\Lambda^0(\bar{\Lambda}^0)$	2182	$\textcircled{\Lambda^0} \longrightarrow \textcircled{p} + \textcircled{\pi^-}$	$\textcircled{p} + \textcircled{\pi^-}$	2.6×10^{-10}
SIGMA PARTICLES	$\Sigma^0(\bar{\Sigma}^0)$	2326	$\textcircled{\Sigma^0} \longrightarrow \textcircled{\Lambda^0} + \textcircled{\gamma}$	$\textcircled{\Lambda^0} + \textcircled{\gamma}$	$< 10^{-11} (\sim 10^{-18}?)$
	$\Sigma^+(\bar{\Sigma}^+)$	2328	$\textcircled{\Sigma^+} \longrightarrow \textcircled{p} + \textcircled{\pi^0}$	$\textcircled{p} + \textcircled{\pi^0}$	$\sim 8 \times 10^{-11}$
XI PARTICLES	$\Sigma^-(\bar{\Sigma}^-)$	2342	$\textcircled{\Sigma^-} \longrightarrow \textcircled{n} + \textcircled{\pi^-}$	$\textcircled{n} + \textcircled{\pi^-}$	1.7×10^{-10}
	$\Xi^-(\bar{\Xi}^-)$	2585	$\textcircled{\Xi^-} \longrightarrow \textcircled{\Lambda^0} + \textcircled{\pi^-}$	$\textcircled{\Lambda^0} + \textcircled{\pi^-}$	$\sim 10^{-10}$
	$\Xi^0(\bar{\Xi}^0)$?	$\textcircled{\Xi^0} \longrightarrow \textcircled{\Lambda^0} + \textcircled{\pi^0}$	$\textcircled{\Lambda^0} + \textcircled{\pi^0}$?

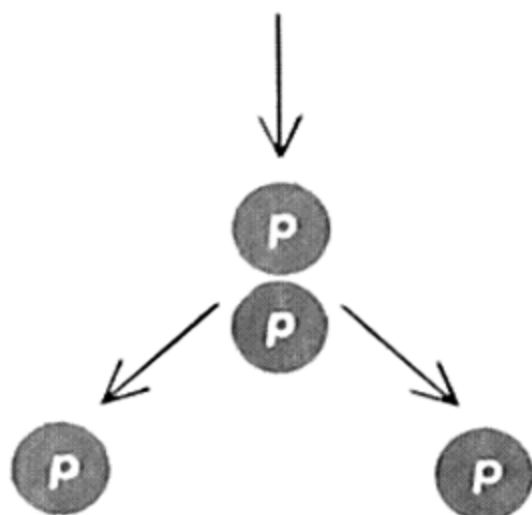
10^{-10} second. On the time-scale of the strong interactions this is incredibly slow. If we magnify 10^{-23} second to a full second, 3×10^{-10} second becomes one million years! It may still seem that the "long" time interval of 3×10^{-10} second is in fact rather short and perhaps difficult to measure. But this is not so. Suppose the lambda particle is traveling with a speed one-third that of light, actually a rather low value in typical experiments. In its brief span of life it will travel three centimeters, quite a macroscopic and measurable distance. The lambda-decay process, typical of the weak interactions, needs all the time in the world.

From such observation we deduce that the weak interactions have only 10^{-14} the strength of the strong interactions. This is not an entirely precise statement, because our theories are crude. But the statement is striking enough. Between the strong interaction and the weak there is an enormous gap in strength.

Reference to the familiar electromagnetic interaction shows just how weak the weak interactions are. We think of electromagnetic forces as weak compared to the nuclear forces; they are in fact somewhat less than 10^{-2} (to be more exact, .0073) as strong. But the weak interactions are 12 decimal places weaker!

Beta-Decay

The most instructive and the best known of the weak interactions is beta-decay. This is one of the processes, first observed at the turn of the century, by which naturally occurring radioactive elements give evidence of their instability. In a typical beta-radioactivity process, a neutron (neutral particle) in the nucleus spontaneously breaks down into a proton and an electron. Since the negatively charged electron flies off and the positively charged proton remains



LOW-ENERGY COLLISION of two protons leaves the particles unchanged. They carom off each other like billiard balls.

bound in the daughter nucleus, the atom is transformed to a species with one more unit of charge. The nature of the process was not clarified, of course, until some years after the radioactivity was first observed.

If beta-decay can happen in some nuclei, why not in all? R. P. Feynman of the California Institute of Technology has given one answer: If matters stood otherwise, we would not be here to ask the question! A somewhat less theological and more fruitful explanation invokes the inviolate law of conservation of energy. For most nuclei beta-decay is prohibited by the law because the mass (a form of energy, in accord with the famous prescription $e=mc^2$) of the nucleus is smaller than the sum of the masses of the electron and the potential daughter nucleus. In a radioactive nucleus the situation is reversed, and the inherent instability of the neutron is allowed to manifest itself. The mass of the neutron itself exceeds the mass of proton and electron by the energy equivalent of 780,000 volts—a very tiny amount as these matters go. As a result the neutron ought to resist beta-decay. It does. Its lifetime is about 18 minutes, by far the longest of any unstable elementary particle on record.

But we are not yet done with energy conservation. The excess 780,000 volts ought to appear in the energy of motion, that is, the kinetic energy, of the decay products. Careful measurement of neutron decay, however, demonstrates beyond any doubt that the proton and electron almost always carry off less kinetic energy (though never more), the amount varying from one event to another. When this situation was faced for the first time, the alternatives seemed grim. Physicists either had to accept a breakdown of the law of energy conservation, or they had to suppose the existence of a new and unseen particle. Such a particle, emitted along with the proton and the electron in the disintegration of the neutron, could save the central pillar of physics by carrying off the missing energy. This was in the early 1930s, when the introduction of a new particle was not the casual matter it is today. Nevertheless, after only the briefest vacillation, physicists chose the second alternative. Enrico Fermi, following ideas outlined originally by Wolfgang Pauli, spelled out the properties of the missing particle, which he named the neutrino.

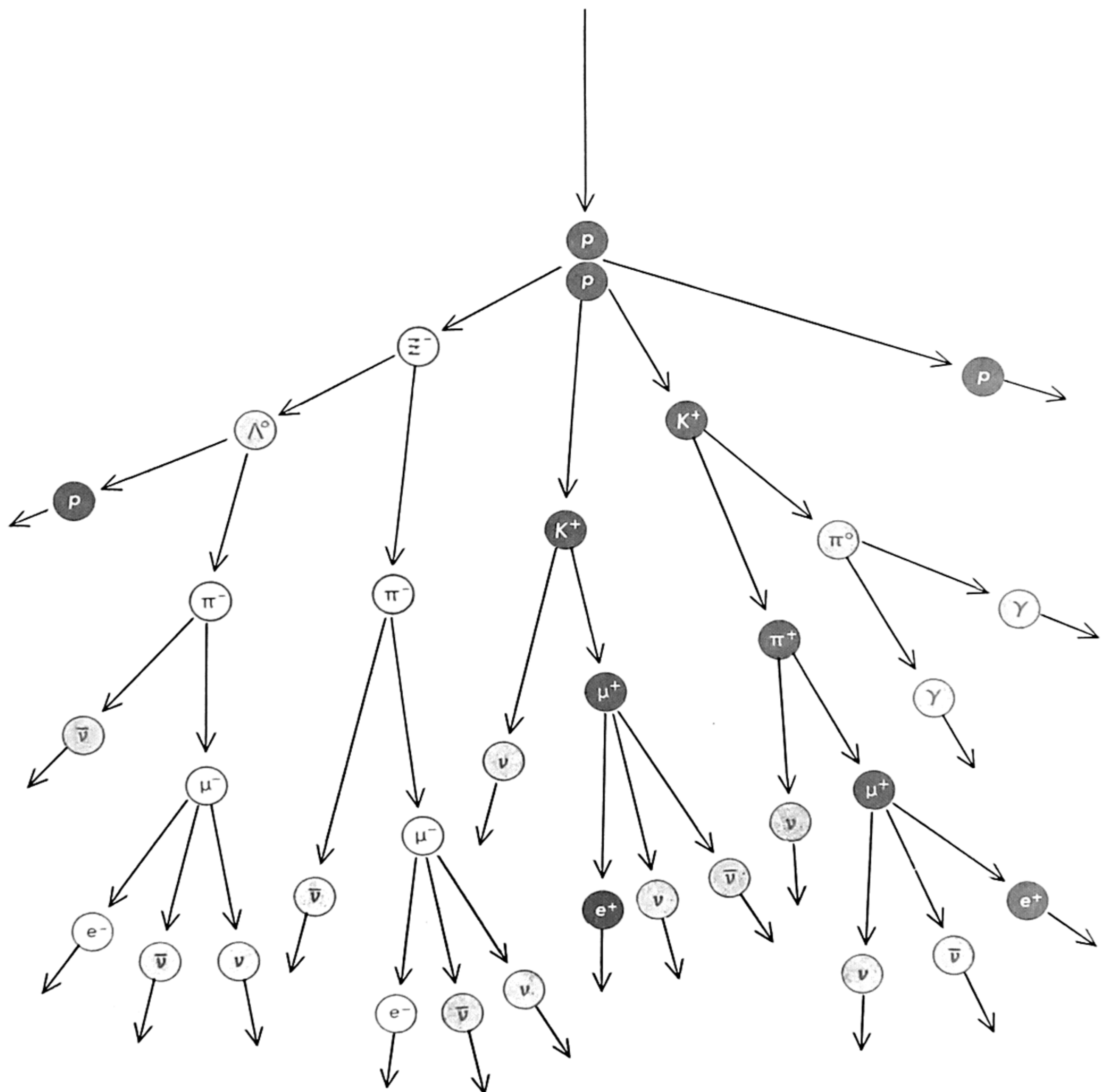
No wonder the neutrino had been unseen! To balance the electrical charge in neutron beta-decay, it must be a neutral

particle. Moreover, from subsequent beta-decay experiments we know that the neutrino has essentially no mass; indeed, there is every theoretical reason to suppose the mass is precisely zero. And the neutrino is essentially unstoppable and unreacting. The probability, for example, that a neutrino passing through the solid earth would be slowed down or do anything other than act as though it were in a vacuum is but one part in one million million. Only in recent years has it become possible to detect the neutrino directly, to catch it in the act of "doing something." The experimental effort required is heroic [see "The Neutrino," by Philip Morrison; SCIENTIFIC AMERICAN Offprint 230].

In many ways the neutrino is the strangest particle of all. It carries off the "missing" energy in beta-decay; it also accounts for the missing linear momentum and angular momentum—quantities which, like energy, satisfy basic conservation laws. Without the neutrino we could not have both beta-decay and the conservation laws. Yet once it is produced the neutrino heads off to the ends of the universe and, so far as we know, does nothing more. Neutrinos from the sun and other stars (and from nuclear reactors) course through matter but are in no active sense part of it.

The New Particles

The conservation of energy plays the same pivotal role in the strong interactions. Suppose a proton is accelerated in a cyclotron and is then allowed to collide with another proton. What happens? At low energies a simple deflection takes place: the target particle is kicked into motion and the incident particle bounces off in some new direction. Just as when billiard balls collide, the final particles are the same as the initial ones and the energy with which they rebound equals the energy with which they collided. But go to higher energies, say 350 million electron volts. Competing reactions now set in. Sometimes the final products consist of a proton plus a neutron plus a new particle: the positive pi meson. It has been created *de novo* out of the energy made available by the collision. One of the protons has transformed into a neutron and a pi meson, revealing not only a new particle but a new interaction. At even higher collision energies one finds still other reactions taking place, other more massive particles being created in various combinations: K mesons, lambda



HIGH-ENERGY COLLISION of two protons produces a shower of particles. One of the original protons darts away unchanged (upper right). The other gives rise to a negative xi particle and

two positive K mesons. These particles, being unstable, then decay into other particles, some of which in turn decay. At the end only stable particles remain, some 20 of them in the event depicted here.

particles, sigma particles, xi particles.

Much as these new particles may complicate the picture of matter, their production in the strong interactions serves to balance the energy account. The same is true of the other particles produced in collisions of these particles and in the processes, mostly weak interactions, by which they decay. The complexity of possible reactions and the profusion of particles are indicated by the diagram

above, which depicts a typical sequence of reactions initiated by a very high-energy collision of two protons.

The proton-proton reaction suggests another ground for distinction between the strong and weak interactions, related to their respective intrinsic strengths. Usually one of the protons utilizes the collision energy to change into a neutron and a positive pi meson, via a strong interaction. But quite an-

other weak reaction ought sometimes to take place. The proton should also be able to avail itself of collision energy to go through a beta transformation into a neutron plus a positron plus a neutrino. This reaction is permitted by the conservation of energy (also by the conservation of linear and angular momentum and by the conservation of charge, the positive charge of the proton being carried off by the positron, the anti-

particle of the electron). No doubt this reaction sometimes occurs when protons collide. But its intrinsic strength is so incredibly small that the process is never in fact observed. It happens, presumably, only once in 10^{14} collisions or so. This is a general situation. Whenever circumstances are such that both strong and weak interactions are possible, the strong overwhelmingly predominate. The weak processes reveal themselves only where the energy supply is such that the conservation of energy forbids the strong interaction:

The weak interactions, however, are more common than this restriction might suggest. Consult the table of elementary particles on pages 394 and 395. Other than the photon, neutrino, electron and proton (and their antiparticles) every one of them is unstable. Only two of the unstable particles decay rapidly via strong processes which happen to be allowed by energy conservation. The remaining particles decay by way of weak interactions. The weak-decay lifetimes vary over a wide range, as one sees in

the table. But most, and perhaps all, of this variation has to do with the differences in energy release. When analyzed in terms of intrinsic interaction strengths, the variations all but disappear. They are always of the same order: 10^{-14} the strength of the strong interactions.

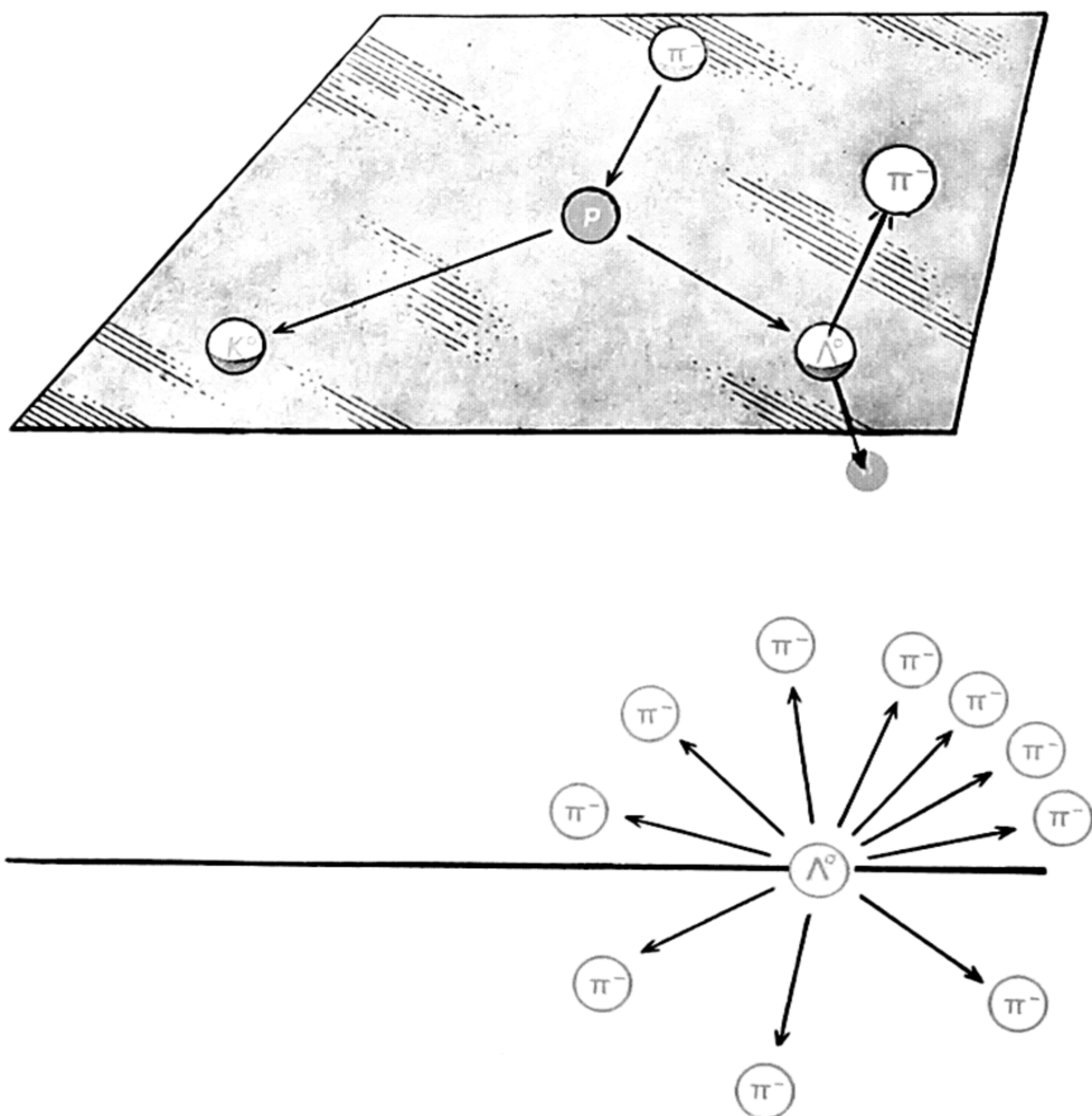
Nobody doubts that this pattern reflects some deep unity underlying the whole set of weak interactions, but as yet we have no clear idea what this unity might be. One thing at any rate is clear. Weakness and strength, as in human affairs, are not intrinsic properties of the individual particles, but only of reactions among particles. There is but one exception: the neutrino enters only into weak processes; it is all weak.

The Overthrow of Parity

What is the use of the weak interactions anyhow? Until recently they attracted little attention; they served to clear away the debris of high-energy collisions, the province of the strong interactions, removing the unstable prod-

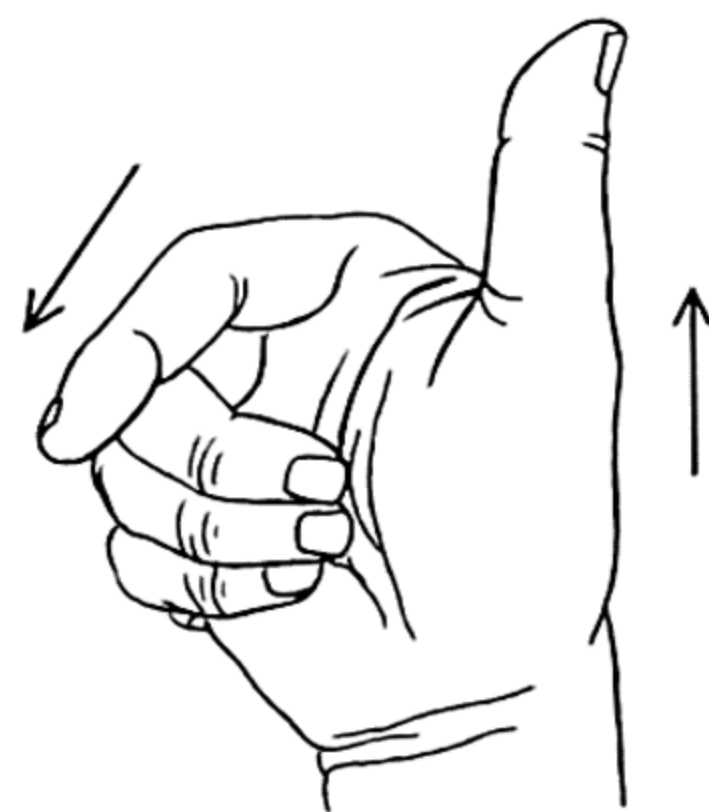
ucts through leisurely decay processes—vital work perhaps, but a bit dull. Physicists nevertheless had to face the fact of these processes. When they began to look more closely, they found a paradox. The law of conservation of parity was being violated.

The puzzle had to do with the K meson. Sometimes it decays into two pi mesons, and sometimes into three. But this should be impossible; with all the conservation laws in force, the K meson would be permitted to decay in one manner or the other, but not both. Perhaps, it was at first thought, there are two kinds of K meson, each accounting for one of the two decay modes. It was easier to accept the existence of a new particle than to face the violation of a conservation law. But no evidence of any other differences between the two particles could be found; they really looked like one and the same. The situation grew untenable. Then, in 1956, Tsung Dao Lee of Columbia University and Chen Ning Yang of the Institute for Advanced Study drew the correct infer-



RIGHT-HANDED CONVENTION is applied to the production and decay of the lambda particle (Λ^0). In the drawing at top a negative pi meson (π^-) strikes a proton (p), giving rise to the lambda particle and a neutral K meson (K^0). The lambda particle

then decays into a proton and a negative pi meson. In the right-handed convention the negative pi meson tends to go off "above" the plane formed by path of the lambda particle and that of the neutral K meson. This preference is suggested by drawing at bottom.



ence. They reminded their colleagues that the parity principle, tested and found trustworthy throughout the whole of the better-explored realm of strong interactions, had never in fact been tested for the weak interactions. Maybe the rule broke down in this other-worldly class. They proposed concrete tests that could decide the issue.

The story is by now a familiar one. Chien Shiung Wu of Columbia University and a group of collaborators at the National Bureau of Standards undertook a set of difficult and beautiful experiments on the beta-decay of polarized nuclei of cobalt 60. In January of 1957 they announced the overthrow of the parity principle.

Shortly thereafter groups at Columbia and the University of Chicago reported the same result from experiments involving the decay of π and μ mesons; parity is not conserved here either. Later still a group at the University of California studied the decay of the charged K meson into a μ meson and a neutrino; once again parity was over-

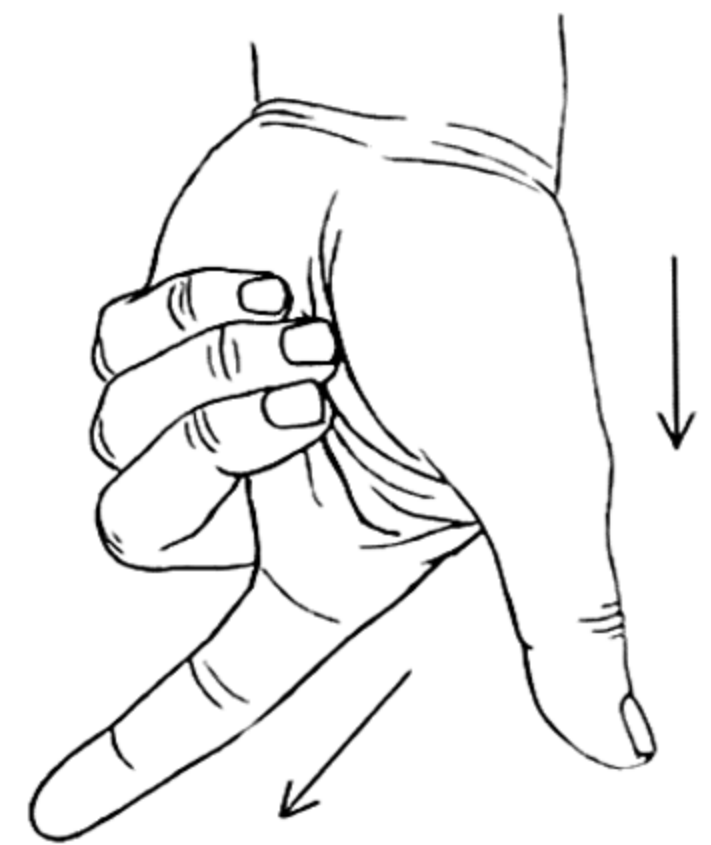
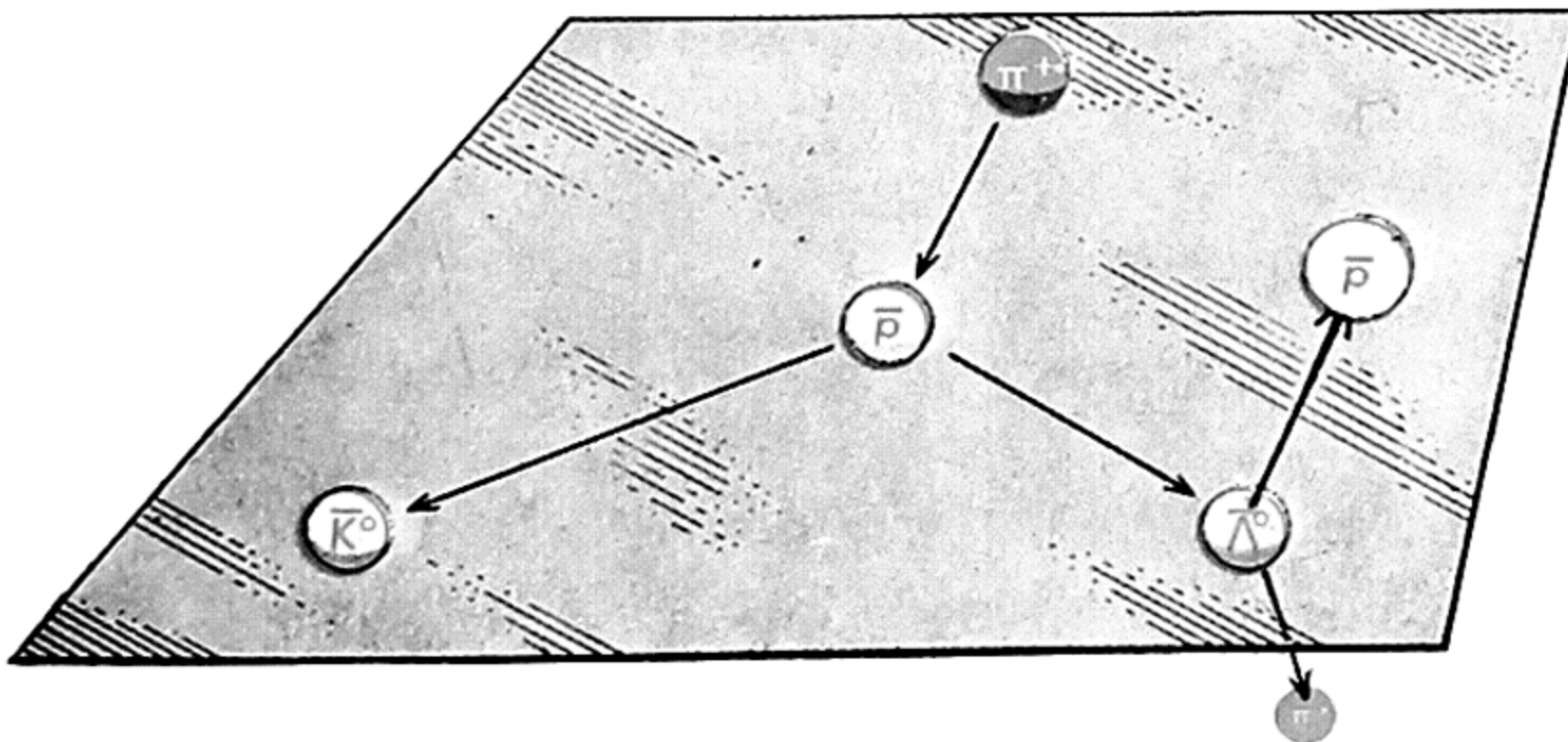
thrown. Other workers proceeded to confirm the downfall of the battered principle in a great variety of nuclear beta-decay transformations and in the beta-decay of the free neutron itself. It was like a bandwagon at a political convention. The weak interactions were now at the center of the stage. In 1957 Miss Wu attended the International Conference on High-Energy Physics, a conclave devoted mainly to the strong interactions; "I am here," she could say, "on the strength of the weak interactions."

The Lambda Particle

Some questions still remained. All of the processes for which the parity principle had been directly shown to be invalid have in common a special feature: in each there is at least one neutrino among the decay products. Could it be that parity violation is peculiar only to the neutrino, a particle which has borne other heavy burdens in the past? To be sure, the neutrino is not involved

in the paradoxical decay of the K meson which had provoked all of these startling developments, but the evidence for parity violation in this case is indirect. A direct experimental test in at least one non-neutrino process was required before one could feel safe in attributing parity violation to weak interactions generally. The lambda particle offered an opportunity for such a test, and experiments were undertaken by groups working at the Brookhaven National Laboratory and at the University of California. By the summer of 1957 the results became known: here, too, parity is overthrown. The precise details of the manner and degree of parity violation in the various weak processes remain to be established. The breakdown of parity, however, now seems to be a quite general property of weak interactions, a further reflection of the deep unity that underlies them and sets them apart from the strong processes.

The parity principle is simple and classical. It asserts that the laws of nature do not distinguish between right



LEFT-HANDED CONVENTION is applied to the production and decay of the anti-lambda particle ($\bar{\Lambda}^0$). In the drawing at top a positive pi meson (π^+) strikes an antiproton (\bar{p}), giving rise to the anti-lambda particle and a neutral anti-K meson (\bar{K}^0). The anti-

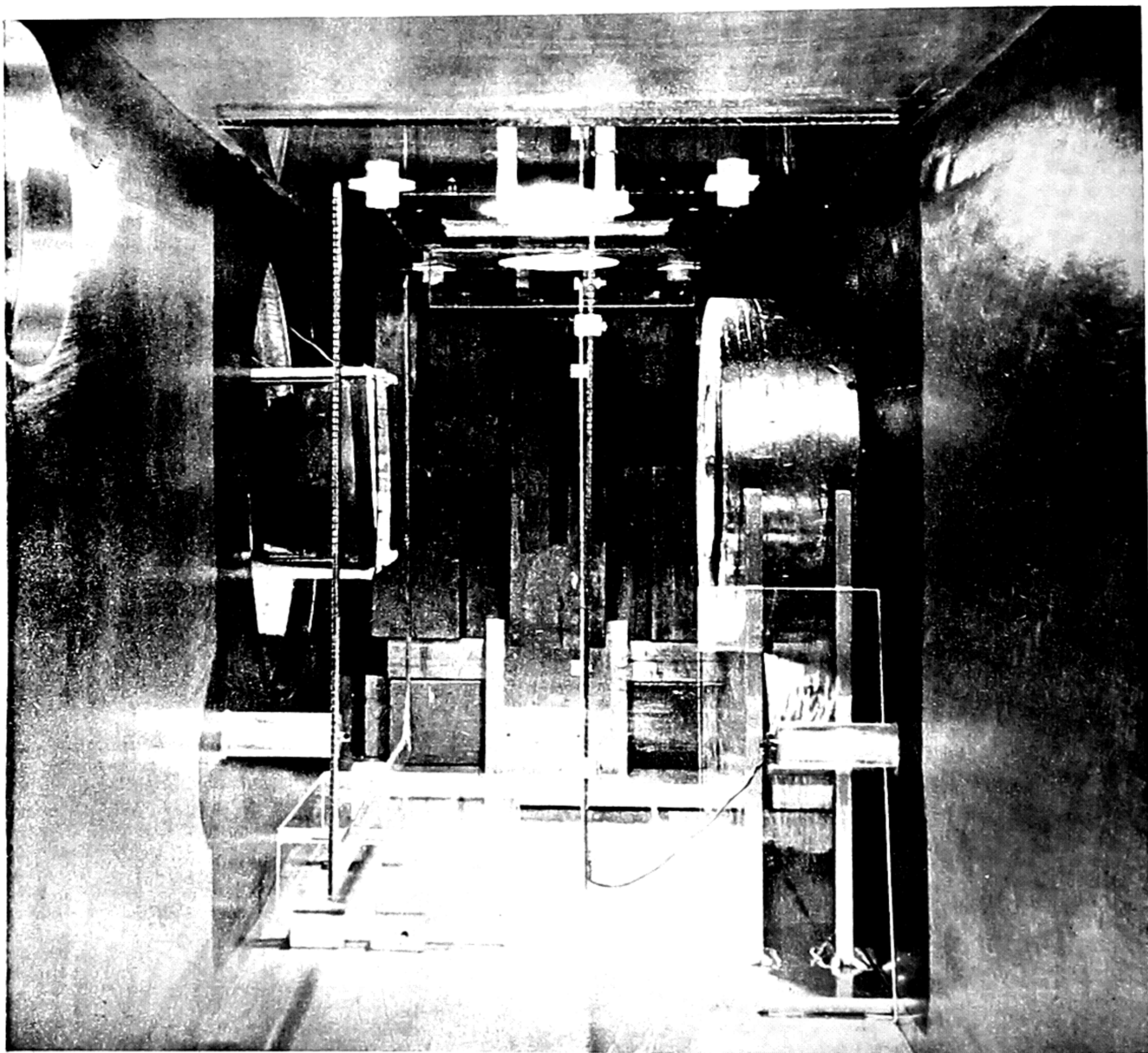
lambda particle then decays into an antiproton and a positive pi meson. If combined symmetry holds, in the left-handed convention the positive pi meson tends to go off "above" plane formed by the path of the anti-lambda particle and that of the neutral anti-K meson

and left. You may think of examples that contradict this assertion; our hearts, for example, are nearly always on our left sides. But this is merely a fact of nature, not a law. In the domain of atomic and nuclear physics no intrinsic difference between right and left had ever been found. As sometimes happens in such situations, the powerful and fruitful principle of left-right symmetry had as-

sumed the status of a self-evident truth, though voices were raised from time to time to remind us that few truths in physics are self-evident.

Let us look more closely at the implications of parity in the context of the production and decay of the lambda particle. In the tests that have actually been carried out, protons are bombarded by a beam of high-energy (one-billion-elec-

tron-volt) negative pi mesons. Among the competing reactions that take place, the one that concerns us is the process: pi meson plus proton yields lambda particle plus K meson. The lambda particle emerges, and we observe its decay, usually into a negative pi meson and a proton [see illustration on page 398]. The line of flight of the original bombarding pi meson and the line of flight of



EXPERIMENTAL APPARATUS in this photograph was employed at the Argonne National Laboratory to demonstrate that parity was not conserved in the beta-decay of the neutron, another weak interaction process. Neutrons from a nuclear reactor were directed at a magnetized cobalt "mirror" which reflected only those neutrons with their "north" and "south" poles aligned in one way. This sent a beam of polarized neutrons to the vacuum tank shown in the photograph through the vertical slot in the background. As the

neutrons passed down the tank a small fraction of them decayed into protons and electrons. The boxlike object at the left side of the tank is a proton counter. Within the cylinder at right is an electron counter. By this means the experimenters were able to show that the electrons tended to go off in a preferred direction, thus violating the conservation of parity. The experimenters were M. T. Burgy, V. E. Krohn, T. B. Novey and G. R. Ringo of the Argonne National Laboratory and V. L. Telegdi of the University of Chicago.

the emerging lambda particle together define a plane, just as any two intersecting lines have done ever since Euclid. Now let us ask about the pi meson that is produced in the disintegration of the lambda particle. It can come off in any direction. But consider in particular only the two directions which are perpendicular to the plane: the "above" direction and the "below" direction. How do we decide which is above and which below? The words have no meaning until we make an arbitrary convention. Let us adopt such a convention.

Take your right hand and direct its index finger along the line of flight of the incident pi meson, and its middle finger along the line of flight of the lambda particle. Let us then say that your thumb is pointing "above" the plane. This fixes the meaning of "above," and hence of "below." If instead you were to take your left hand and follow exactly the same instructions, "above" and "below" would be interchanged.

Now the parity principle asserts that nature sets up no such arbitrary conventions to distinguish left and right; she does not care which set of definitions you choose. The decay pi mesons should thus emerge with equal probability "above" and "below" the plane. The experiments, however, showed that nature does make such a distinction. It was found that, in the sense of the right-handed convention, the pi mesons come out more often "above" than "below."

Another Symmetry Overthrown

Very soon after the parity principle was called into question, and even before the first tests were carried out, physicists began to worry about another of their beloved symmetry principles: the symmetry between matter and antimatter. This principle, too, had served faithfully in the strong interactions, but it had never been tested for the weak interactions.

The principle in question here is not a simple, geometrical one like the left-right symmetry we have been discussing. It has its origin in the theory of quantum mechanics, and it is somewhat complex in detail. Nonetheless the principle can be described in a rather straightforward manner. The whole thing started with a beautiful mathematical theory of the electron that P. A. M. Dirac of the University of Cambridge developed 30 years ago. The theory immediately proved to be enormously successful in accounting for the detailed behavior of electrons in atoms. But in another direction it led to a startling result: the pre-

diction of a new particle, the positron. This antiparticle to the electron, with the same mass but opposite charge, was discovered several years later. The same theory also predicted the existence of antiprotons and antineutrons, particles that have been found only in more recent experiments. Dirac's ideas have since been generalized to cover all of particle physics. We now believe that to every particle there corresponds a distinct antiparticle, the only exceptions being the photon and the neutral pi meson, where particle and antiparticle are one and the same. In fact, even the neutrino has an antineutrino; it is the antineutrino which is emitted in neutron beta-decay. The Dirac theory, and its subsequent generalizations, had been constructed in such a way as to treat particles and antiparticles on an equal footing; that is, every reaction involving particles should have a corresponding reaction involving antiparticles, and the two sets of reactions should proceed in the same way in every detail. Understandably this notion of a basic symmetry between matter and antimatter assumed the status of a firm principle of physics. It threatened to become a self-evident truth.

Let us again consider lambda-particle decay and ask what the symmetry between matter and antimatter entails. The diagram on page 398 depicts the production and the subsequent decay of a lambda particle, and shows how these processes violate the parity principle. Now turn to the diagram on page 399. Here we deal with the corresponding antiparticle processes. Suppose that the initial energy of the anti-pi meson as well as the angle of its flight with the line of flight of the lambda antiparticle are the same as for the particle reactions shown in the diagram on page 398. The two diagrams thus differ only in that particles have been replaced by antiparticles. Now in the particle reaction we know that the decay pi mesons come off preferentially in the "above" direction, as defined in the right-handed convention. This preference reflects the breakdown of the parity principle. Be that as it may, the principle of symmetry between matter and antimatter requires that the same preference for "above," defined by the same right-handed convention, should hold in the antiparticle reaction. In other words, particle and antiparticle should both violate parity in the same way. If it were otherwise, there would be an intrinsic distinction between matter and antimatter, which the symmetry principle forbids.

As a matter of fact our hypothetical

antiparticle experiment has not yet been carried out, though it no doubt will be in due course. However, the symmetry of matter and antimatter not only relates their reactions to one another; it also has certain internal consequences for either kind of reaction taken alone. The analysis is rather technical. What it comes to in the case of the particle processes is this. If matter-antimatter symmetry holds, then the asymmetry between "above" and "below" cannot exceed a certain calculable amount. In the experiments that have been carried out, this upper limit is exceeded by a wide margin. This means that the decay pi meson in the antiparticle reaction comes off preferentially in the "below" direction, as defined by the right-handed convention. Similar tests have been made for beta-decay, and for pi-, mu- and K-meson decay, with the same result. We conclude that symmetry between matter and antimatter does not obtain in these processes or, presumably, in weak interactions generally.

Two cherished symmetry principles have now been undermined by the weak interactions: left versus right, matter versus antimatter. The question is: Can anything of these symmetries yet be saved? There is one possibility. What happens if we simultaneously interchange right and left, matter and antimatter? Perhaps this combined symmetry survives, even in the world of the weak interactions. The diagram on page 399 shows what this would entail. In the particle reactions it is a fact that the decay pi mesons prefer "above," as defined by a right-handed convention. If the combined symmetry is valid, then in the antiparticle reactions the decay pi mesons should have the same preference for "above," but the "above" direction is now defined by a left-handed convention.

Once again it is not actually necessary to go to the antiparticle experiments to test combined symmetry. The principle has internal consequences for purely "particle" reactions, and these are accessible to test. Highly accurate and difficult beta-decay experiments now under way should soon decide the matter with high reliability. At the moment the outlook for the principle of combined symmetry is favorable.

Suppose the principle holds true. Then it would be no violation of the laws of nature if somewhere in the universe a Tsung Dao anti-Lee and a Chen Ning anti-Yang are presently searching their hearts for the reason why. Their hearts, of course, would be on their right sides.

The Author

S. B. TREIMAN is a theoretical physicist at Princeton University. A Chicagoan, he "drifted to mathematics in high school and, on bum advice, started college as a chemical engineer." This career was interrupted by World War II and service as a Navy radar technician. "I switched to physics after the war," he reports, "for no strong reason except that during a whole year in the Philippines my only reading material besides detective stories (of which I read at least 100) was, of all things, a popular exposition of relativity, atomic physics and all the other wonders of nature—badly written as I later learned but fascinating at the time." Treiman took

his undergraduate and graduate degree at the University of Chicago, then joined the Princeton staff, where he is now an associate professor.

Bibliography

AN INTRODUCTION TO THE PHYSICS OF THE NEW PARTICLES. C. Franzinetti and G. Morpurgo in *Il Nuovo Cimento*, Supplement VI, Numero 2, pages 469-800; 1957.

K-MESONS AND HYPERONS: THEIR STRONG AND WEAK INTERACTIONS. R. H. Dalitz in *Reports on Progress in Physics*, Vol. 20, pages 163-303. The Physical Society of London, 1957.

THE PHYSICS OF ELEMENTARY PARTICLES. J. D. Jackson. Princeton University Press, 1958.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

RADIATION BELTS AROUND THE EARTH

by James A. Van Allen

Instruments borne aloft by artificial satellites and lunar probes indicate that our planet is encircled by two zones of high-energy particles, against which space travelers will have to be shielded.

So far, the most interesting and least expected result of man's exploration of the immediate vicinity of the earth is the discovery that our planet is ringed by a region—to be exact, two regions—of high-energy radiation extending many thousands of miles into space. The discovery is of course troubling to astronauts; somehow the human body will have to be shielded from this radiation, even on a rapid transit through the region. But geophysicists, astrophysicists, solar astronomers and cosmic-ray physicists are enthralled by the fresh implications of these findings. The configuration of the region and the radiation it contains bespeak a major physical phenomenon involving cosmic rays and solar corpuscles in the vicinity of the earth. This enormous reservoir of charged particles plays a still-unexplained role as middleman in the interaction of earth and sun which is reflected in magnetic storms; in the airglow and in the beautiful displays of the aurora.

The story of the investigation goes back to 1952 and 1953, before any of us could think realistically about the use of earth satellites to explore the environment of the earth. Parties from our laboratory at the State University of Iowa spent the summers of those years aboard Coast Guard and naval vessels, cruising along a 1,500-mile line from the waters of Baffin Bay, near the magnetic pole in the far northwestern corner of Greenland, southward to the North Atlantic off the coast of Newfoundland. Along the way we launched a series of rocket-

carrying balloons—"rockoons." (The balloon lifts a small rocket to an altitude of 12 to 15 miles, whence the rocket carries a modest payload of instruments to a height of 60 to 70 miles.) Our objective was to develop a profile of the cosmic-ray intensities at high altitudes and latitudes, and thus to learn the nature of the low-energy cosmic rays which at lower altitudes and latitudes are deflected by the earth's magnetic field or absorbed in the atmosphere.

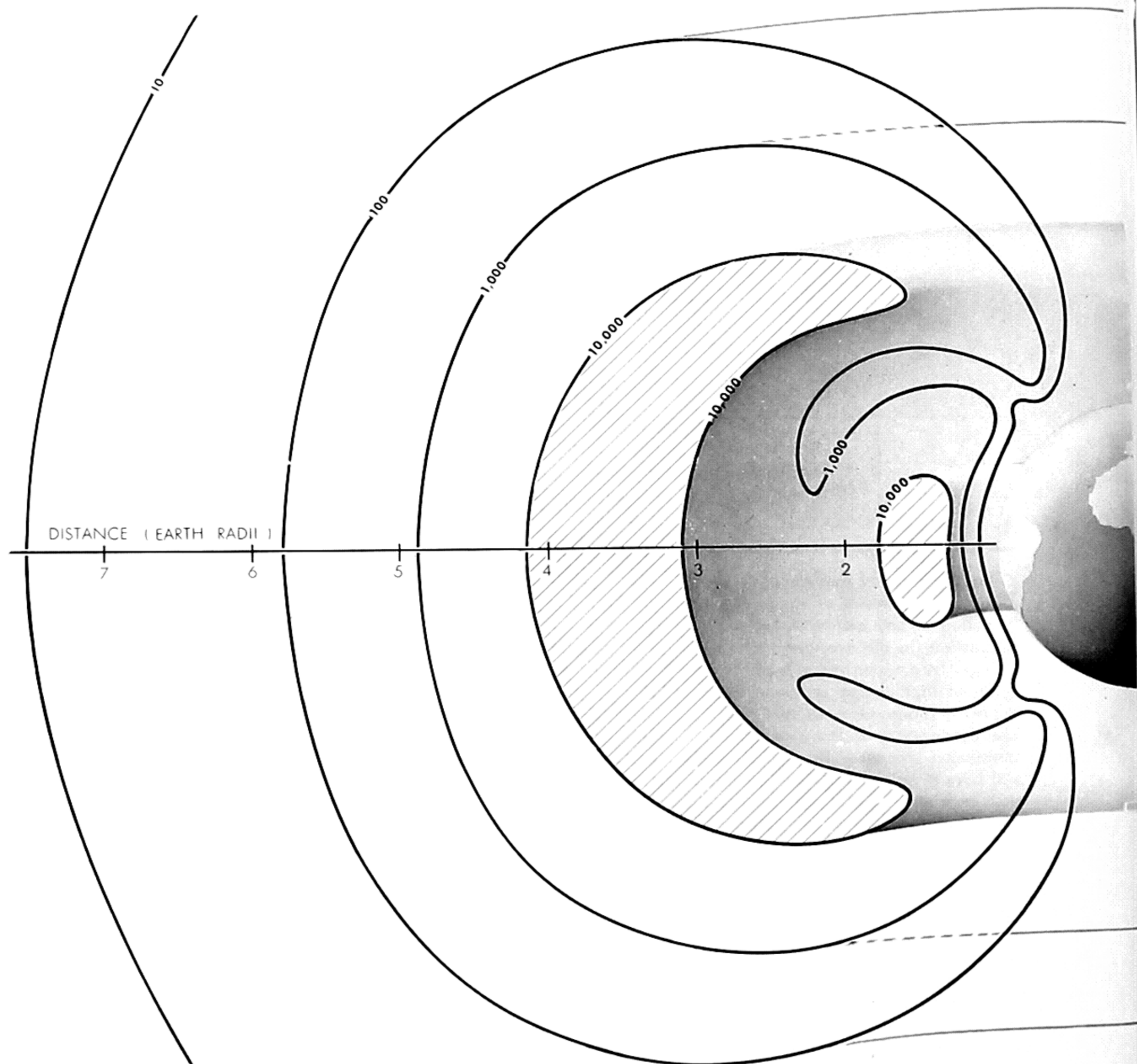
Most of the readings radioed down from the rockets were in accord with plausible expectations. Two rockoons sent aloft in 1953, however, provided us with a puzzle. Launched near Newfoundland by Melvin Gottlieb and Leslie Meredith, they encountered a zone of radiation beginning at an altitude of 30 miles that was far stronger than we had expected. At first we were uneasy about the proper operation of our instruments. But critical examination of the data convinced us that we had unquestionably encountered something new in the upper atmosphere.

Significantly these measurements were made in the northern auroral zone. In this zone, which forms a ring some 23 degrees south of the north geomagnetic pole, the incidence of visible auroras reaches its maximum. Since rockets fired north and south of the zone had revealed nothing unusual, we speculated that the strong radiation played some part in the aurora. Showers of particles from the sun, it was thought, come plunging into the atmosphere along magnetic lines of

force and set off these displays [see "Aurora and Airglow," by C. T. Elvey and Franklin E. Roach; SCIENTIFIC AMERICAN, September, 1955]. But the theory underlying this explanation did not explain satisfactorily why the aurora and the high-intensity radiation we had detected should occur in the auroral zone and not in the vicinity of the geomagnetic pole itself. Nor could it account for the high energies required to carry the solar particles through the atmosphere to such relatively low altitudes.

The mystery deepened when we found in later studies that the radiation persists almost continuously in the zone above 30 miles, irrespective of visible auroral displays and other known high-altitude disturbances. More discriminating detectors established that the radiation contains large numbers of electrons. Our original observations had detected X-rays only; now it turned out that the X-rays had been generated by the impact of electrons on the skin of the instrument package (as if it had been the "target" in an X-ray tube) and on the sparse atoms of the upper atmosphere itself. Sydney Chapman and Gordon Little at the University of Alaska suggested that such a process might well account for the attenuation of radio signals in the lower ionosphere of the auroral zones.

The International Geophysical Year gave us our first opportunity to investigate the "auroral soft radiation" on a more comprehensive scale. During the



STRUCTURE OF RADIATION BELTS revealed by contours of radiation intensity (*black lines*) is shown schematically by shading

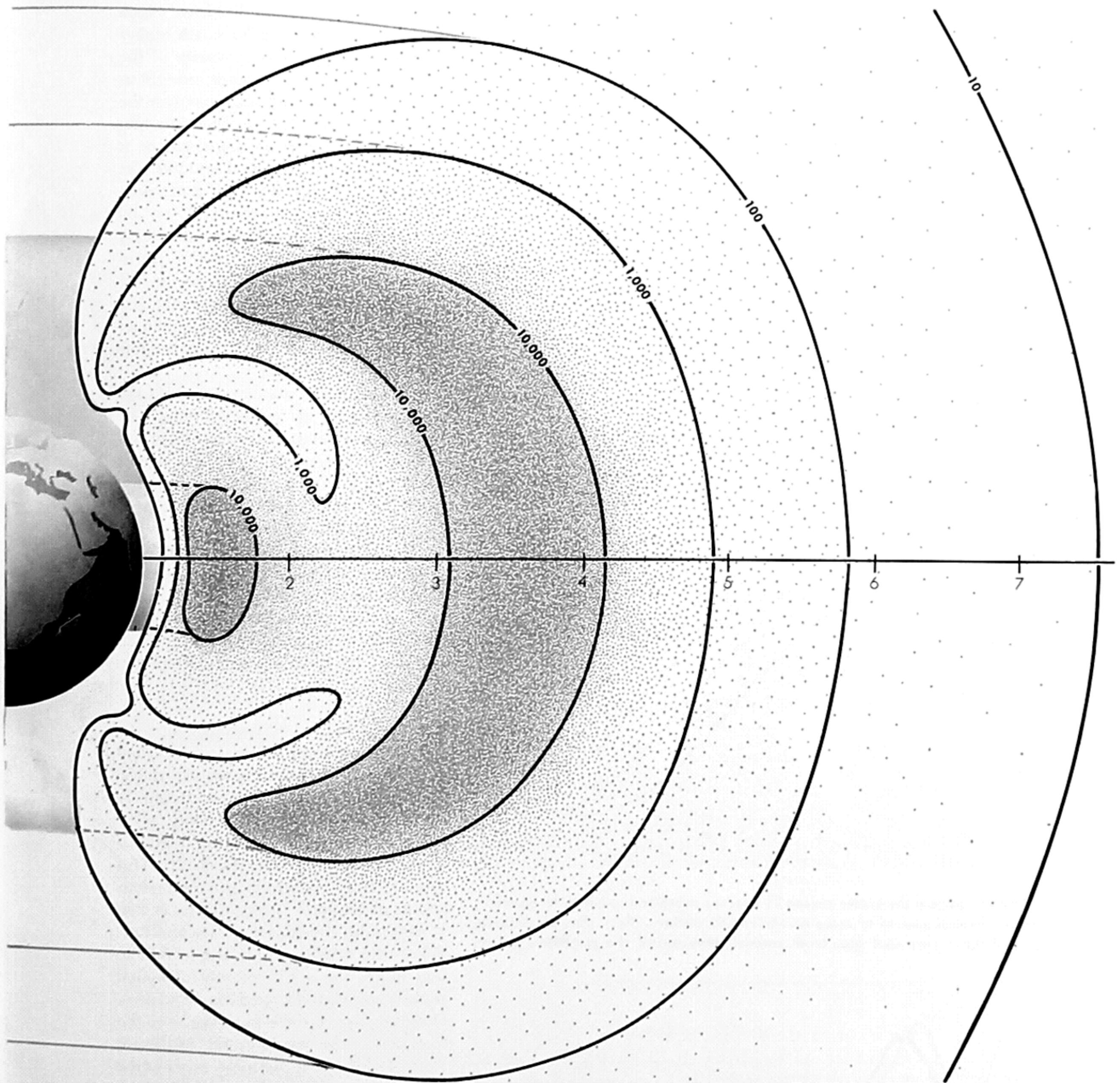
(*left*); dots (*right*) suggest distribution of particles in the two belts. Contour numbers give counts per second; horizontal scale

summer and fall of 1957 Laurence Cahill and I launched a number of rockoons off the coast of Greenland and also got off one successful flight in Antarctica. The latter flight established that the radiation exists in the southern as well as the northern auroral zone. In February, 1958, Carl McIlwain fired a series of two-stage rockets through visible auroras above Fort Churchill in Canada, and discovered that the radiation includes

energetic protons (hydrogen nuclei) as well as electrons.

Meanwhile all of us had been pushing a new development that greatly expanded the possibilities for high-altitude research. During the summer of 1955 the President and other Government authorities were finally persuaded that it might be feasible to place artificial satellites in orbit, and authorized an I. G. Y. project for this purpose. In January,

1956, a long-standing group of high-altitude experimentalists, called the Rocket and Satellite Research Panel, held a symposium to consider how the satellites could be most fruitfully employed. At that meeting our group proposed two projects. One was to put a satellite into an orbit nearly pole-to-pole to survey the auroral radiation in both the north and south auroral zones. Such orbits, however, did not appear to be



shows distance in earth radii (about 4,000 miles) from the center of the earth. Particles in the inner belt may originate with the

radioactive decay of neutrons liberated in the upper atmosphere by cosmic rays; those in the outer belt probably originate in the sun.

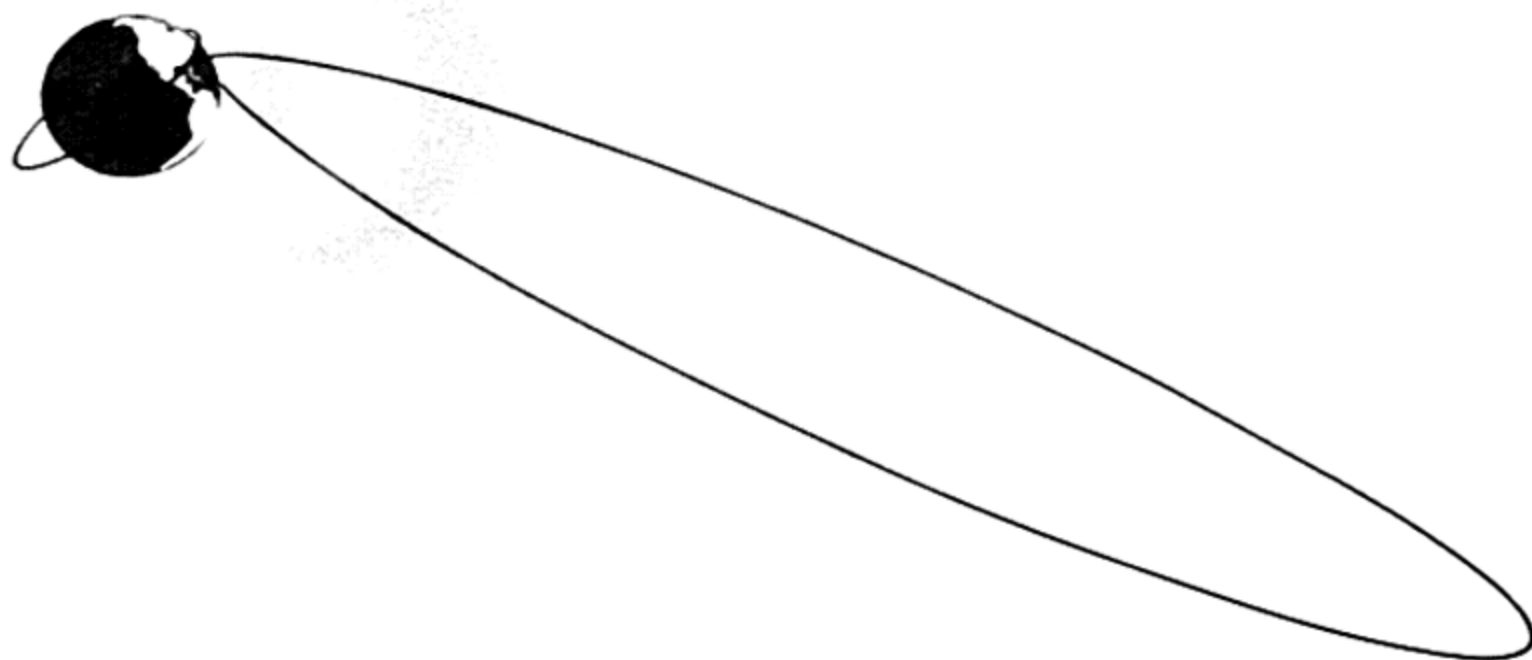
technically feasible in the immediate future. For the time being we were forced to abandon the use of a satellite to probe farther into the auroral soft radiation. We also suggested that a satellite orbiting over the lower latitudes of the earth might usefully be employed in a comprehensive survey of cosmic-ray intensities over those regions. This project was adopted, and we were authorized to prepare suitable experimental

apparatus [see "The Artificial Satellite as a Research Instrument," by James A. Van Allen; *SCIENTIFIC AMERICAN*, November, 1956]. It was planned to place this apparatus on one of the early Vanguard vehicles.

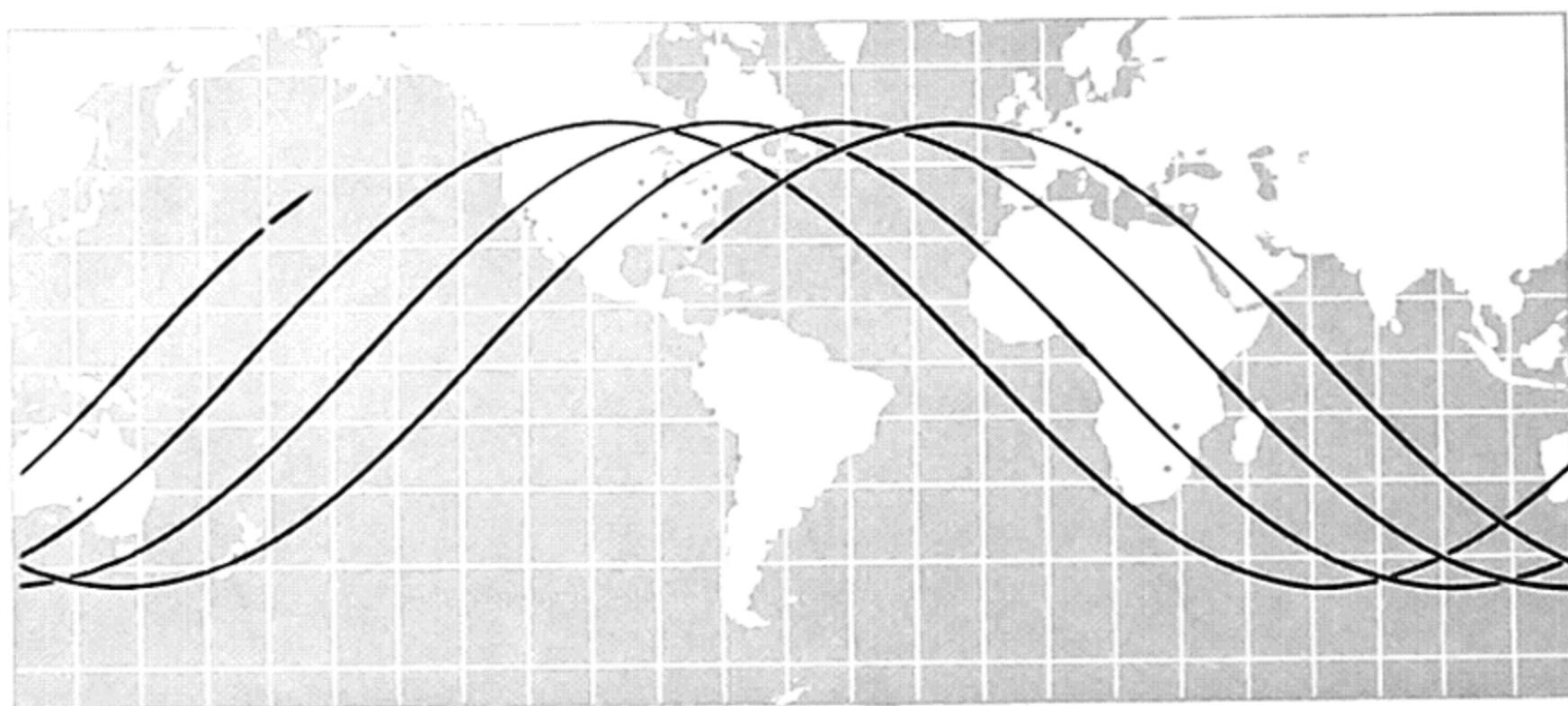
The difficulties and failures of the Vanguard are now history. Sputnik I stimulated some high government officials to accept a proposal that a number of us had been urging for more than

a year: to use the proven Jupiter C rocket as a satellite-launching vehicle. As a result on January 31, 1958, Explorer I went into orbit carrying our simple cosmic-ray detector and a radio to broadcast its readings.

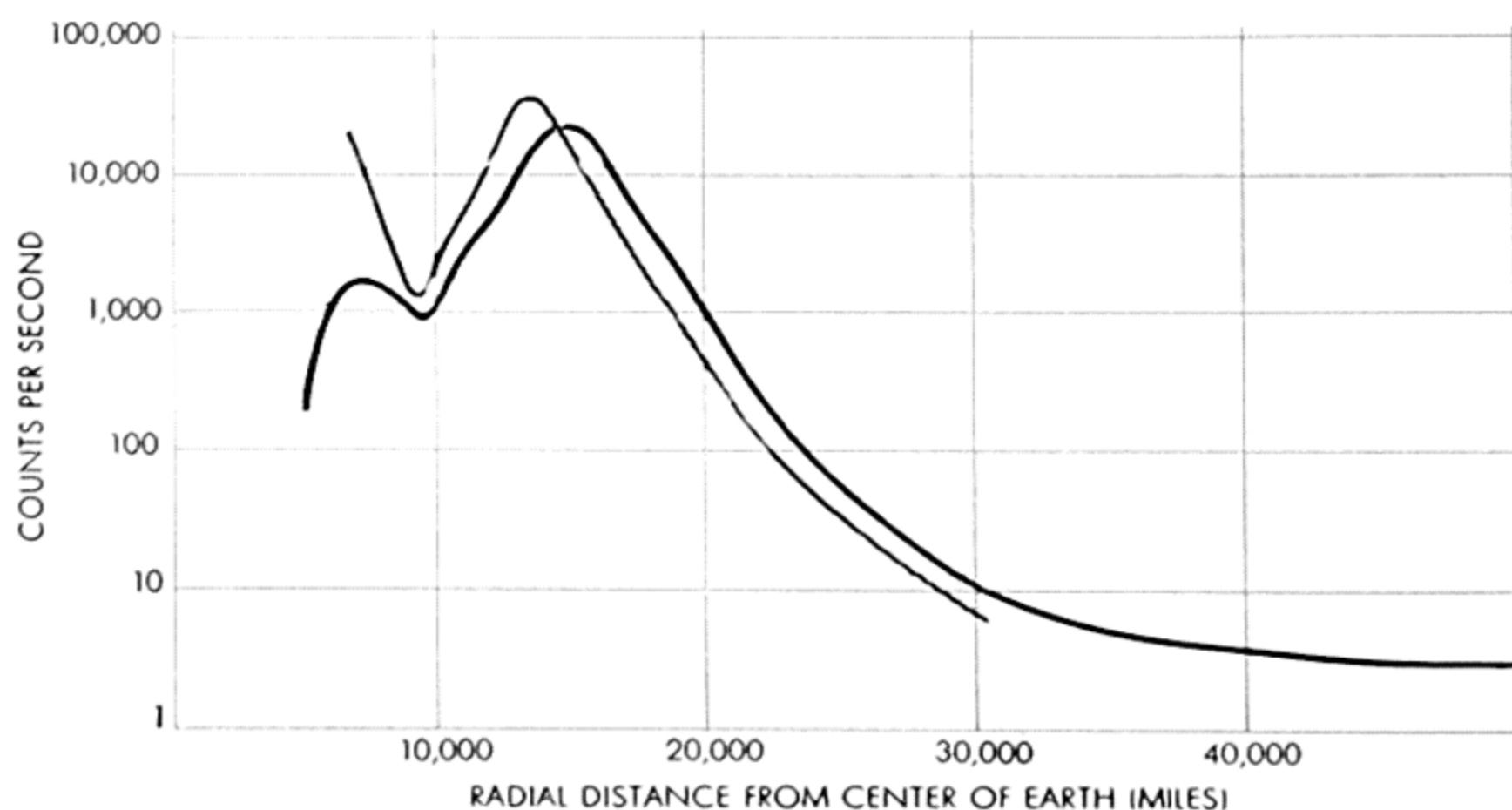
In the first reports from stations located in the U. S. the intensity of radiation increased with altitude along the expected curve. Several weeks later, however, we began to get tapes from stations in



EXPLORER IV AND PIONEER III gave the first detailed picture of the radiation belts. The Explorer IV satellite (*short ellipse*) monitored radiation levels for nearly two months at altitudes up to 1,300 miles. The Pioneer III lunar probe (*long ellipse*) provided data out to 65,000 miles. Its orbit is shown distorted because of the earth's rotation during flight.



EXPLORER IV ORBIT covered the entire region 51 degrees north and south of the equator; the black curve shows a small part of its trace on the earth's surface. More than 25 observation stations (*colored dots*) recorded data from several thousand of the satellite's passes.



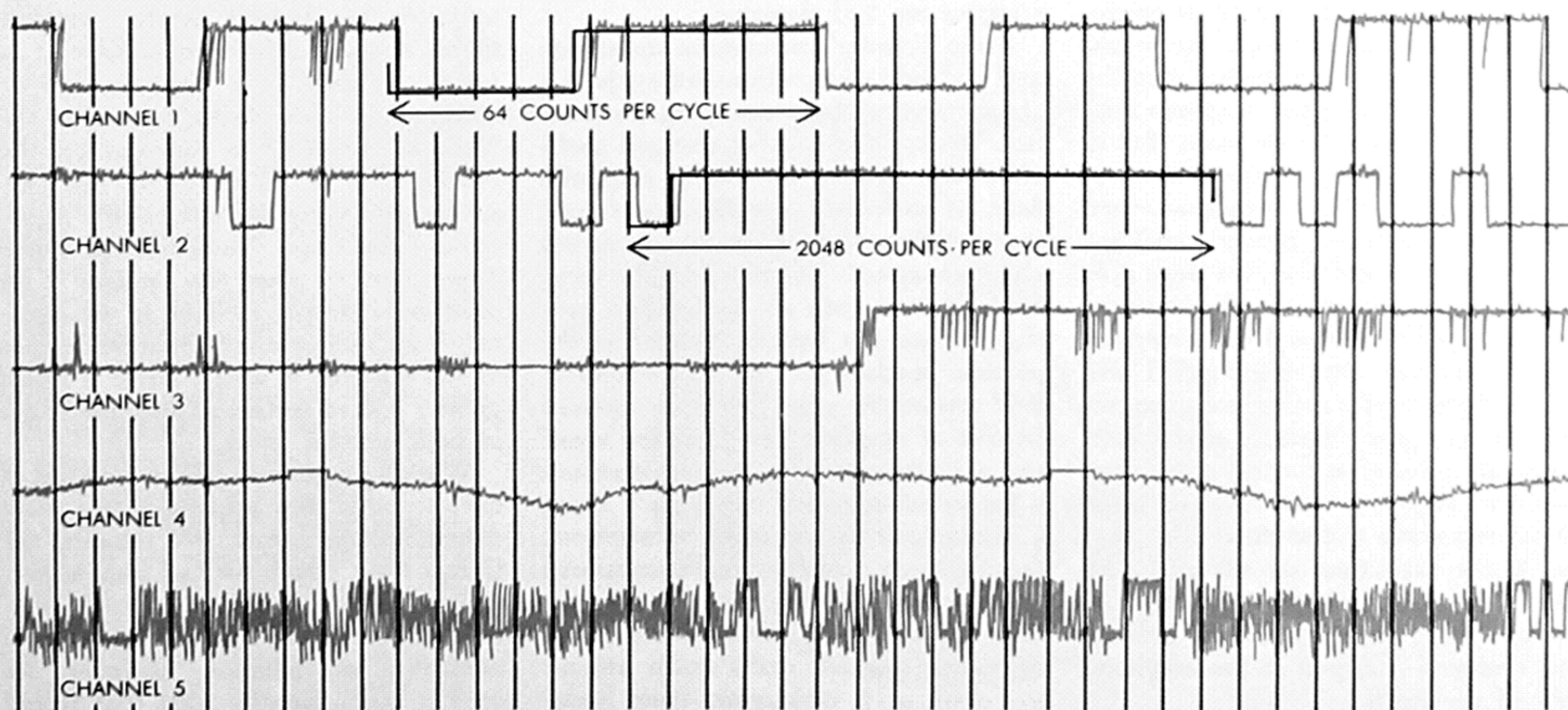
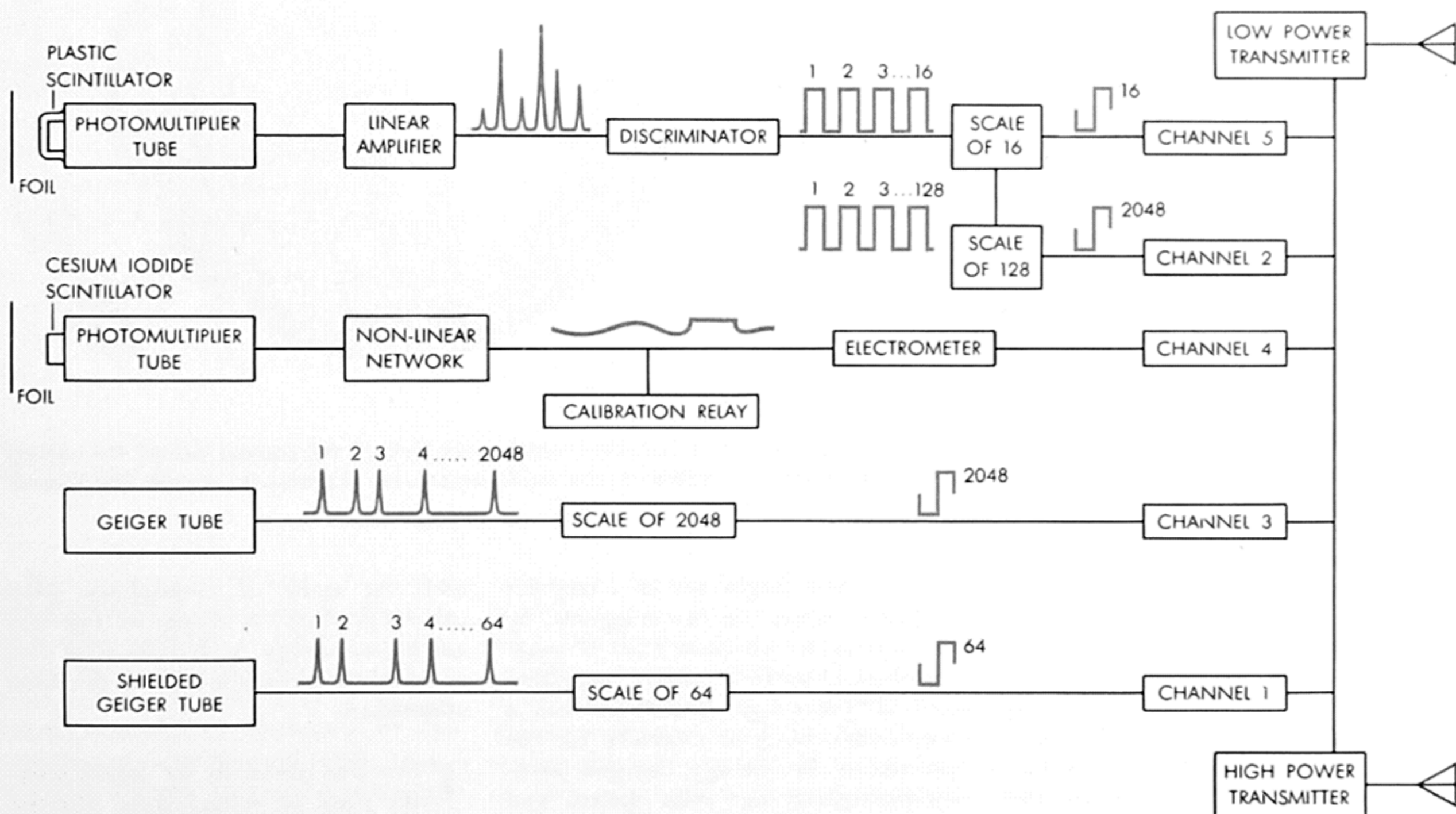
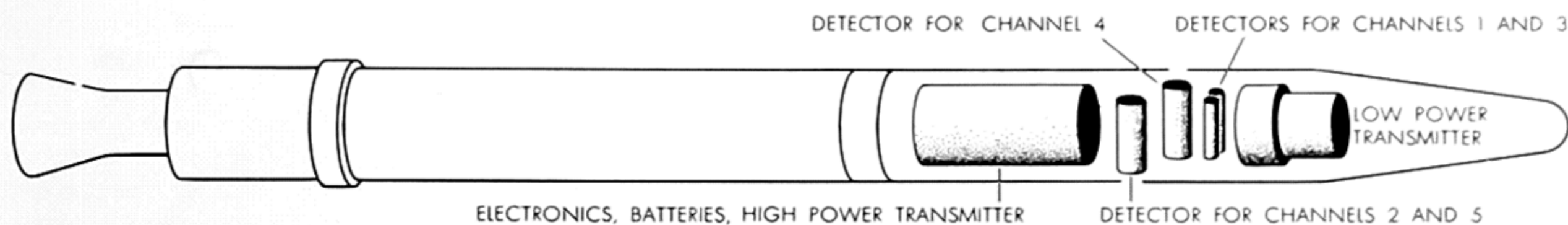
PIONEER III DATA gave the first confirmation of two distinct rings of particles. Counting rates on both the outbound (*black curve*) and the inbound (*gray curve*) legs of the flight showed two peaks. The two curves differ because they cover different sections of the belts.

South America and South Africa which gave us counting rates for much higher altitudes, due to the eccentricity of the satellite's orbit. These records brought us a new surprise. At high altitudes over the equatorial region the apparent counting rate was very low; in some passes it dropped to zero for several minutes. Yet at lower altitudes the rate had quite "reasonable" values—from 30 to 50 counts a second. Again we were uneasy about the trustworthiness of the instruments. The only alternative seemed to be that cosmic rays do not strike the uppermost layers of the atmosphere over the tropics, and we were quite unable to accept this conclusion.

Our uneasiness was increased by the incompleteness of our early data. The Explorer I apparatus broadcast its observations continuously, but its signals could be picked up only intermittently, when the satellite came within range of a ground station. Our original apparatus, designed and developed by George Ludwig for the Vanguard satellites, included a magnetic-tape recorder which could store its observations for a complete orbit around the earth and then report them in a "burst" on radio command from the ground.

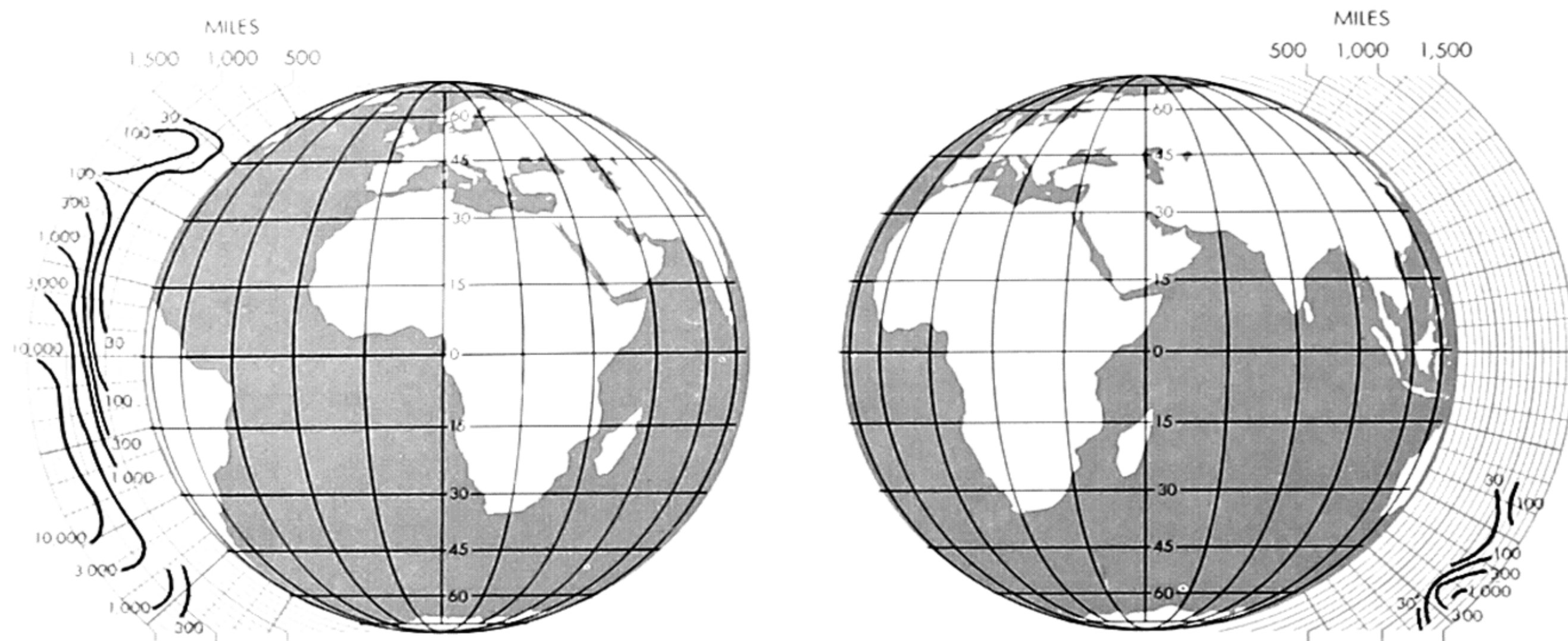
By early February, working with the Jet Propulsion Laboratory, we had converted this apparatus for use in the Explorer II satellite. The first attempt to get it into orbit failed. A second rocket placed Explorer III, carrying identical apparatus, in orbit on March 26. This satellite fully confirmed the anomalous results of Explorer I. At altitudes of 200 to 300 miles the counting rate was low. When the satellite went out to 500 to 600 miles, the apparent rate ascended rapidly and then dropped almost to zero. One day, as we were puzzling over the first tapes from Explorer III, McIlwain suggested the first plausible explanation for their peculiar readings. He had just been calibrating his rocket instruments, and called our attention to something that we all knew but had temporarily forgotten: A sufficiently high level of radiation can jam the counter and send the apparent counting rate to zero. We had discovered an enormously high level of radiation, not a lack of it. As Ernest Ray, a member of our group, inaccurately but graphically exclaimed: "Space is radioactive!"

During the next two months Explorer III produced a large number of playback records, every one of which showed the same effect. At low altitudes the counting rate was reasonably attributable to



EXPLORER IV INSTRUMENTS were designed to give a detailed picture of the nature and intensity of the radiation. Plastic scintillator counted only charged particles above certain energies; two different scaling factors adapted it to both high and low counting rates. Cesium-iodide scintillator measured the total energy input

rather than individual particles. Shielded and unshielded Geiger tubes could be compared to estimate the penetrability of the radiation. Radio signals suggested by color curves in upper drawing were recorded by ground stations and later played through a multichannel oscillograph to yield records like that shown below.



TWO SETS OF CONTOURS from readings on opposite sides of the earth (left and center) show the northern and southern "horns"

of radiation, which point toward the auroral zones; the contour numbers show radiation intensity in counts per second. The "tipped"

cosmic rays. At higher altitudes—the precise height depended on both latitude and longitude—the count increased to very high values. Up to the points at which the counter jammed, it showed counting rates more than 1,000 times the theoretical expectation for cosmic rays. From the rate of increase and the length of the periods of jamming we judged that the maximum count probably went to several times this level. Since the radiation appeared to resemble the auroral soft radiation, we would not have been surprised to find it in the auroral zone or along the magnetic lines of force that connect these zones. But in the equatorial latitudes these lines of force lie much farther out in space than the altitudes attained by the satellites.

On May 1 of last year we were able to report with confidence to the National Academy of Sciences and the American Physical Society that Explorers I and III had discovered a major new phenomenon: a very great intensity of radiation above altitudes of some 500 miles over the entire region of their traverse, some 34 degrees north and south of the equator. At the same time we advanced the idea that the radiation consists of charged particles—presumably protons and electrons—trapped in the magnetic field of the earth.

We could rule out uncharged particles and gamma and X-rays because they would not be confined by the magnetic field, and so would be observed at lower altitudes. The possibility that the earth's

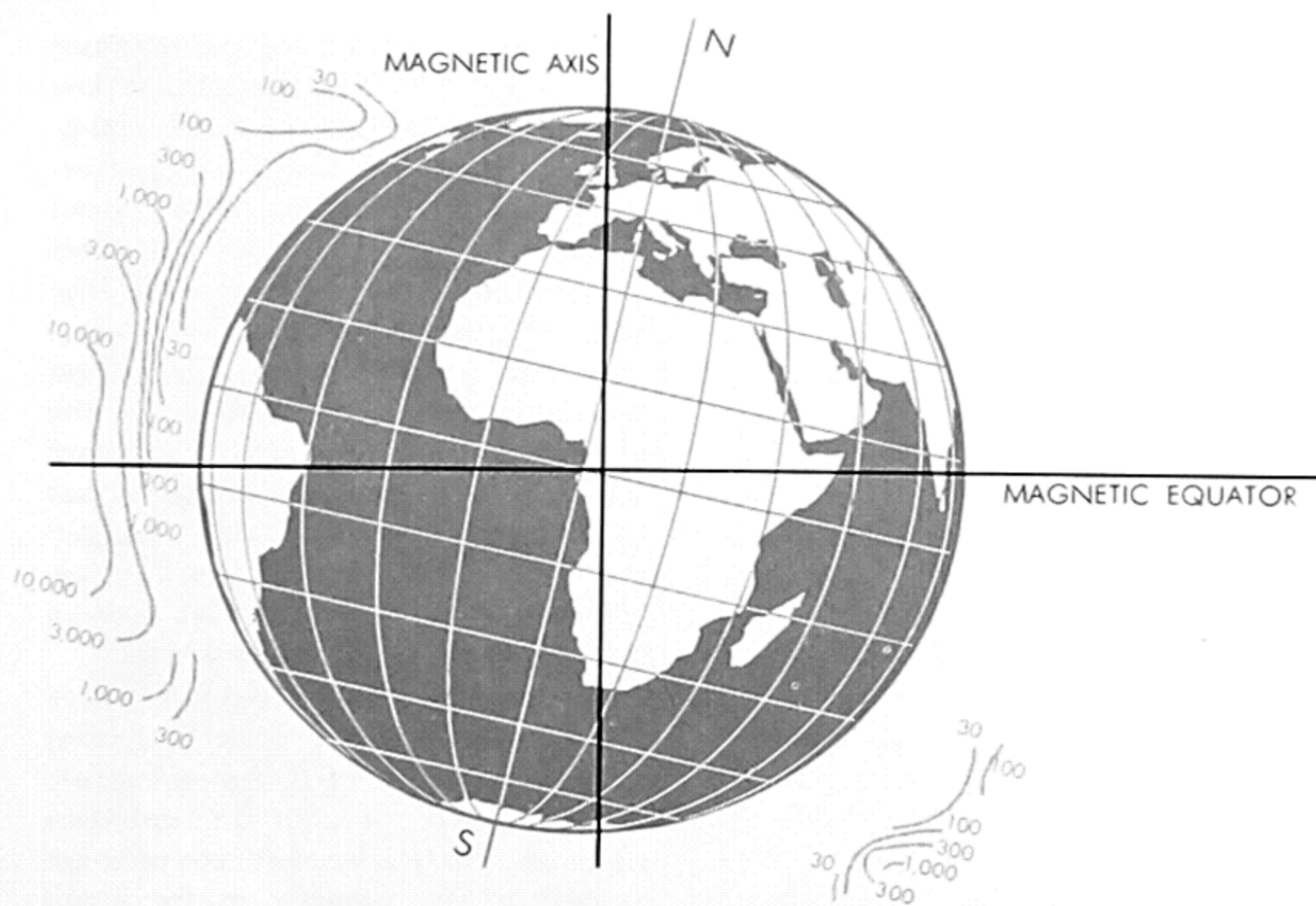
magnetic field might act as a trap for charged particles was first suggested by the Norwegian physicist Carl Störmer in a classical series of papers beginning some 50 years ago, and there was a considerable body of evidence for the existence of low-energy charged particles throughout our solar system and specifically in the vicinity of the earth. But there had been no indication that these particles would possess the high energies we had detected.

From Störmer's theoretical discussion and our own observations we evolved a rough picture of the trapping mechanism. When a fast-moving charged particle is injected into the earth's magnetic field, it describes a corkscrew-shaped trajectory, the center line of which lies along a magnetic line of force. The turns of the helical path are quite open over the equator but become tighter as the particle reaches the stronger magnetic field toward the poles [see illustration at bottom of opposite page]. At the lower end of its trajectory the particle goes into a flat spiral and then winds back along a similar path to the other hemisphere, making the transit from one hemisphere to the other in a second or so. During this time its line of travel shifts slightly, so that the particle drifts slowly around the earth as it corkscrews from hemisphere to hemisphere. An electron drifts from west to east; a proton, in the opposite direction. At each end of its path the particle descends into regions of higher atmospheric density; collisions

with the atoms of atmospheric gases cause it gradually to change its trajectory and to lose energy. After a period of days or weeks the particle is lost into the lower atmosphere.

There was obviously an urgent scientific need to extend these observations with equipment of greater dynamic range and discrimination. In April of 1958 we persuaded several Federal agencies to support further satellite flights of our radiation equipment as an adjunct to the I. G. Y. program, and we received the enthusiastic support of the National Academy of Sciences for the continuation of our work. We also persuaded the Army Ballistic Missile Agency and the Cape Canaveral Air Force Base to try to place the satellite in an orbit more steeply inclined to the equator; at an inclination of about 50 degrees to the equator it would cover a much greater area of earth and skim the edges of both auroral zones.

Working night and day, we set out at once to build new apparatus of a more discriminating nature. We retained the Geiger tube, which we had used in previous satellites, as a basic "simple-minded" detector. To be ready for the highest intensities of radiation, however, we used a much smaller tube that would yield a lower count in a given flux of radiation, and we hooked it into a circuit that would scale down its count by a much larger factor. To obtain a better idea of the penetrability of the radiation



drawing at right shows the essential symmetry of the radiation around the earth's magnetic axis. The structure of the radiation zone was built up from hundreds of observed points.

we shielded a similar Geiger tube with a millimeter of lead. As a more discriminating particle detector we adopted a plastic scintillator and photomultiplier tube to respond to electrons with an energy of more than 650,000 electron volts and to protons of more than 10 million electron volts. Finally we glued a thin cesium-iodide crystal to the window of another photomultiplier tube; the light emitted by the crystal when it was irradiated would measure the over-all input of energy rather than the arrival of individual particles. To keep out light when the crystal faced the sun, we shielded it with thin, opaque nickel foil. A special amplifier gave this detector a large dynamic range extending from about .1 erg per second to 100,000 ergs per second.

Explorer IV carried this apparatus into orbit on July 26, and sent down data for almost two months. Magnetic tapes from some 25 observing stations flowed in steadily from late July to late September; altogether we obtained some 3,600 recorded passes of the satellite. A typical pass was readable for several minutes; some of the best were readable for up to 20 minutes, a large fraction of the time required for the satellite to make a turn around the earth. We are still analyzing this mass of data, but the preliminary results have already proved to be enlightening.

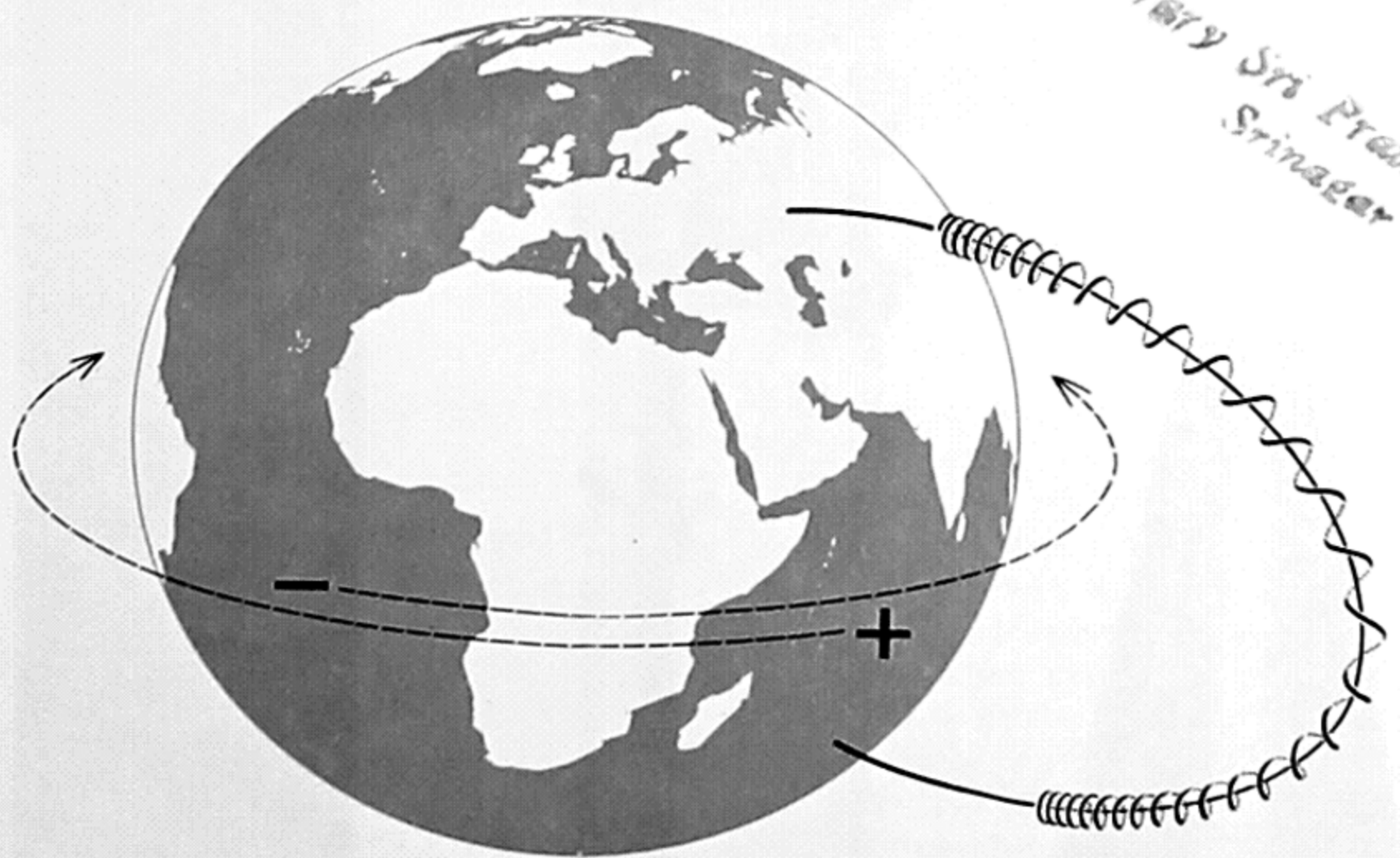
The readings have confirmed our earlier estimates of the maximum levels of radiation. Moreover, we have extended

our observations to more than 50 degrees north and south of the equator and have been able to plot the intensity of the radiation at various latitudes and longitudes for altitudes up to 1,300 miles. The intensity contours follow the shape of the earth in the equatorial region, but as they approach high northern and southern latitudes they swing outward, then inward and sharply outward again to form "horns" reaching down toward

the earth near the auroral zones [see illustrations at the top of these two pages]. The entire picture so far is completely consistent with the magnetic-trapping theory.

It was clear from the contours that Explorers I, III and IV penetrated only the lower portion of the radiation belt. As early as last spring we began to make hypothetical extensions of the observed contours out to a distance of several thousand miles. One of these speculative diagrams showed a single, doughnut-shaped belt of radiation with a ridge around the northern and southern edges of its inner circumference, corresponding to the horns of the contours. Another showed two belts—an outer region with a banana-shaped cross section that extended from the northern to the southern auroral zone and an inner belt over the equator with a bean-shaped cross section [see illustration on pages 404 and 405]. The latter diagram seemed to fit the contours better. In our seminars and after-hour discussions McIlwain held out for the two-belt theory. The rest of us tended to agree with him but preferred to stay with the single "doughnut" because of its simplicity.

To take the question out of the realm of speculation we had to secure measurements through the entire region of radiation. In May, therefore, I arranged to have one of our radiation detectors carried aboard the lunar probes planned for the fall of 1958. On October



TRAPPED PARTICLES spiral rapidly back and forth along a corkscrew-shaped path whose center is a magnetic line of force. At the same time they drift slowly around the earth (broken arrows). Electrons (negative) and protons (positive) drift in opposite directions.

11, 12 and 13 Pioneer I, the first lunar probe, carried our instruments nearly 70,000 miles out from the earth. Though its readings were spotty, they confirmed our belief that the radiation extended outward for many thousands of miles, with its maximum intensity no more than 10,000 miles above the earth.

The next attempted moon shot, Pioneer II, was a fizzle. Pioneer III, however, went off beautifully on December 6. Although this rocket was intended to reach the vicinity of the moon, we were almost as pleased when it failed to do so, for it gave us excellent data on both the upward and downward legs of its flight, cutting through the radiation region for 65,000 miles in two places.

The observations on both legs showed a double peak in intensity [see illustration at bottom of page 406], establishing there are indeed two belts rather than one. The inner belt reaches its peak at about 2,000 miles from the earth, the outer one at about 10,000 miles. Beyond 10,000 miles the radiation intensity diminishes steadily; it disappears almost completely beyond 40,000 miles. The maximum intensity of radiation in each belt is about 25,000 counts per second, equivalent to some 40,000 parti-

cles per square centimeter per second.

Most of us believe that this great reservoir of particles originates largely in the sun. The particles are somehow injected into the earth's magnetic field, where they are deflected into corkscrew trajectories around lines of force and trapped. In this theoretical scheme the radiation belts resemble a sort of leaky bucket, constantly refilled from the sun and draining away into the atmosphere. A particularly large influx of solar particles causes the bucket to "slop over," mainly in the auroral zone, generating visible auroras, magnetic storms and related disturbances. The normal leakage may be responsible for the airglow which faintly illuminates the night sky and may also account for some of the unexplained high temperatures which have been observed in the upper atmosphere.

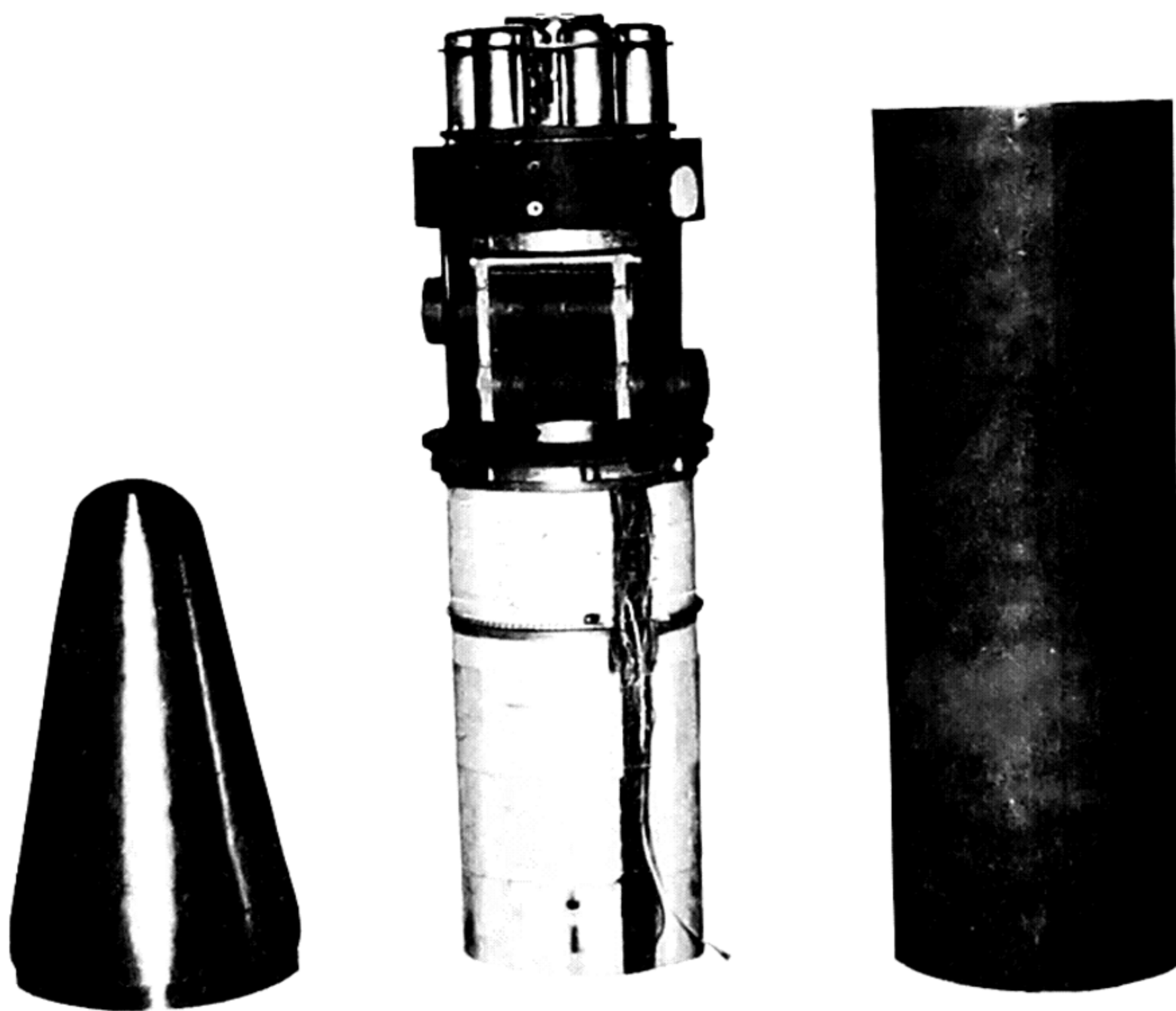
This solar-origin theory, while attractive, presents two problems, neither of which is yet solved. In the first place the energy of many of the particles we have observed is far greater than the presumed energy of solar corpuscles. The kinetic energy of solar corpuscles has not been measured directly, but the time-lag between a solar outburst and the consequent magnetic disturbances

on earth indicates that the particles are slow-moving and thus of relatively low energy. It may be that the earth's magnetic field traps only a high-energy fraction of the particles. Alternatively, some unknown magnetohydrodynamic effect of the earth's field may accelerate the sluggish particles to higher velocities. Some such process in our galaxy has been suggested as responsible for the great energies of cosmic rays. The second problem in the solar-origin theory is that it is difficult to explain how charged particles can get into the earth's magnetic field in the first place. We believe that neither problem is unsolvable.

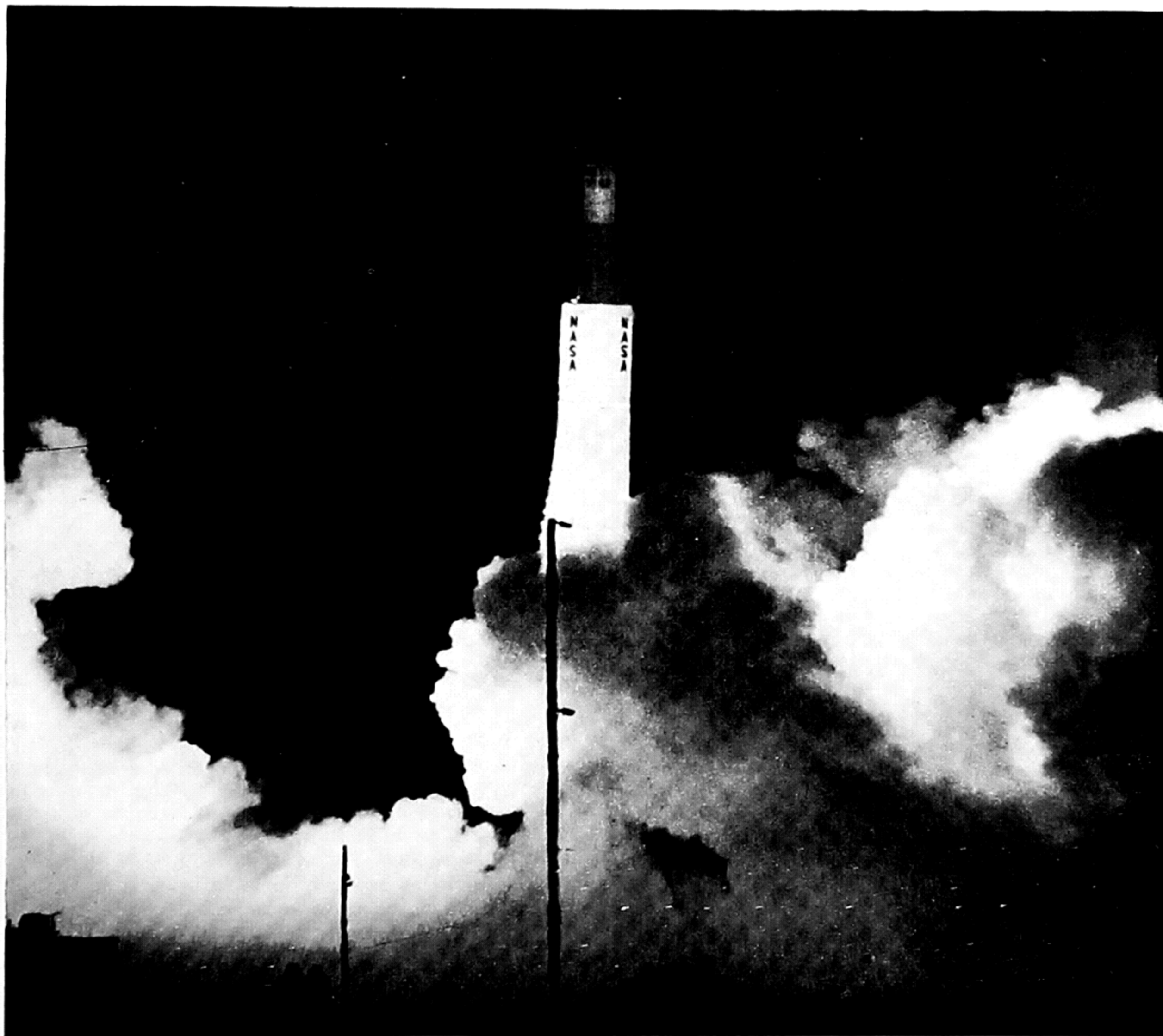
Nicholas Christofilos of the University of California and the Soviet physicist S. N. Vernov have suggested an entirely different theory of how the radiation originates. They note that neutrons are released in large numbers in the earth's upper atmosphere by the impact of cosmic rays. These neutrons, being uncharged, can travel through the magnetic field without deflection. In due course some of them decay there into electrons and protons, which are trapped.

Our group agrees that particle-injection of this sort is going on, and at a rate which can be easily calculated; but we feel for a number of reasons that it cannot be the main source of radiation-belt particles. If we are right in supposing that the radiation belts provide the "reservoir" for the aurora, the neutron hypothesis cannot account for more than one 10,000th of the auroral energy output. Even if the association between the radiation belts and the aurora turns out to be fortuitous, preliminary indications both from our work and from the Russian experience with Sputnik III suggest that most of the particles in the radiation belt have much lower energies than those of particles that would be produced by neutron decay. A full knowledge of the energy distribution of the particles will aid greatly in clarifying their origin.

Neither theory explains why there should be two belts rather than one. It is tempting to combine the two theories and suppose that the inner belt originates with "internal injection"—i.e., neutron-decay products—and the outer one with "external injection" of solar corpuscles. The two-belt configuration may of course be a transitory phenomenon, though the data from Explorer IV and Pioneer III indicate that the separate belts persisted in essentially the same form for at least five months. We should bear in mind, however, that 1958 was a year of great solar activity. Three years



HEAD OF EXPLORER IV includes nose cone (*left*), instrument "payload" (*center*) and protective shell (*right*). Payload includes four detectors, two radio transmitters, batteries and associated electronic circuitry. The outer shell is approximately six inches in diameter.



FOUR-STAGE ROCKET launched the Pioneer III moon probe on December 6, 1958. Though the flight failed to reach the moon, its

outbound leg gave a continuous record of radiation out to 65,000 miles; the inbound leg gave data between 30,000 and 10,000 miles.

from now we may well find a much lower over-all intensity and perhaps a different structure altogether.

In addition to these possible long-term changes, there may be short-term fluctuations in the belts. While we feel sure that the influx and leakage of particles must balance in the long run, a major solar outbreak may temporarily increase the intensity of the radiation many-fold. If we were to detect such fluctuations and were to find that they coincide with solar outbursts on the one hand and with terrestrial magnetic disturbances on the other, we would have a plain lead to the origin of the particles. Before long we hope to launch a satellite

that will monitor radiation levels for at least a year.

Our measurements show that the maximum radiation level as of 1958 is equivalent to between 10 and 100 roentgens per hour, depending on the still-undetermined proportion of protons to electrons. Since a human being exposed for two days to even 10 roentgens would have only an even chance of survival, the radiation belts obviously present an obstacle to space flight. Unless some practical way can be found to shield space-travelers against the effects of the radiation, manned space rockets can best take off through the radiation-free zone over

the poles. A "space station" must orbit below 400 miles or beyond 30,000 miles from the earth. We are now planning a satellite flight that will test the efficacy of various methods of shielding.

The hazard to space-travelers may not end even when they have passed the terrestrial radiation belts. According to present knowledge the other planets of our solar system may have magnetic fields comparable to the earth's and thus may possess radiation belts of their own. The moon, however, probably has no belt, because its magnetic field appears to be feeble. Lunar probes should give us more definite information on this point before long.

The Author

JAMES A. VAN ALLEN heads the physics department at the State University of Iowa. A native Iowan, he received his doctorate from the State University in 1939. During World War II he served as a Navy ordnance and gunnery officer and helped to develop the radio proximity fuze. Later he supervised the development and scientific use of the Aerobee rocket and balloon-launched rockets ("rockoons"). For the past 12 years his group at the University has been studying cosmic rays at high altitudes; Van

Allen directed the cosmic-ray instrumentation of the satellite and space-probe program of the International Geophysical Year.

Bibliography

SPECTRUM OF LOW-RIGIDITY COSMIC RAYS DURING THE SOLAR FLARE OF FEBRUARY 23, 1956. J. A. Van Allen and J. R. Winckler in *The Physical Review*, Vol. 106, No. 5, Second Series, pages 1,072-1,073; June 1, 1957.

THE ARTIFICIAL SATELLITE AS A RESEARCH INSTRUMENT. James A. Van Allen in *Scientific American*, Vol. 195, No. 5; November, 1956.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

THE NATURE OF SOLIDS

by Gregory H. Wannier

The theory explaining the characteristics of substances in the solid state is a relatively recent acquisition of physics. Already this analysis has led to the development of such practical devices as the transistor.

EVERYONE develops in the course of his life a general notion of what a solid body is: that which supports when sat on, which hurts when kicked, which kills when shot. We have also known for a long time certain laws of the behavior of solid bodies: the laws of free fall, of refraction, of elasticity and so on. Yet none of those laws really describes the nature of the solid state. Long after the behavior of gases had been formulated and explained in terms of molecular action, the essential character of the solid state remained a secret. No suggestion of an answer was forthcoming to the question how the same molecules could behave so differently in a solid and in a gas. In fact, the question was not even clearly asked for a

long time, and an unasked question is especially difficult to answer.

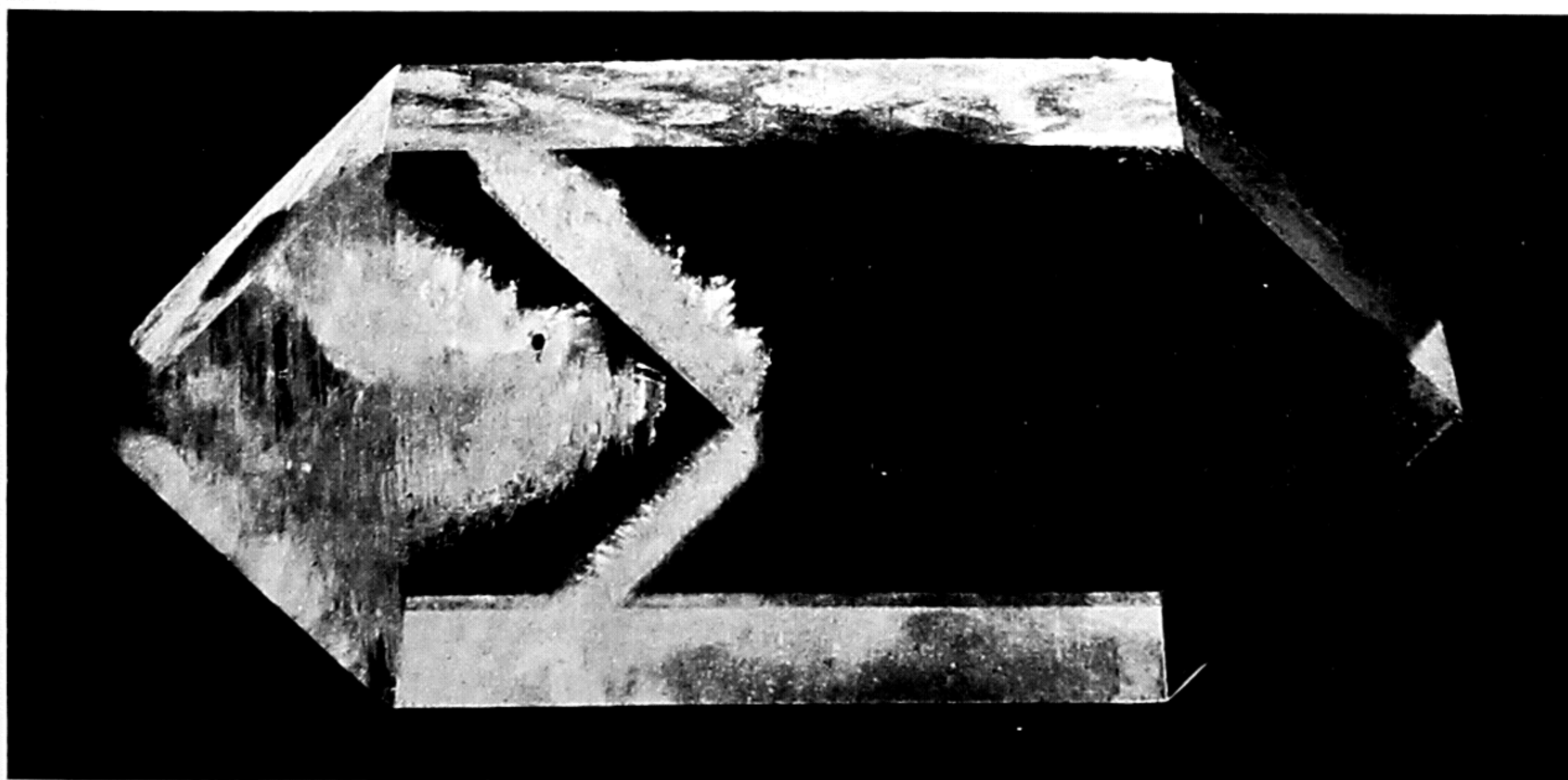
The physics of the solid state is a new science, developed only within the present century. It is often the unusual and startling aspects of a subject that wake it from its slumber. Two such aspects can be singled out as having opened up the physics of the solid state at the turn of the century: the structure of crystals and the conduction of electricity by metals.

Crystals have never suffered from the lack of glamour which so long delayed progress in the physics of the solid state. The travelogues of the ancients are liberally sprinkled with mentions of wondrous gems. The classification of crystals was one of the chief occupations of

Arab scientists. To these early observers, crystals seemed to be exceptional forms of solids. Nineteenth century research disclosed, however, that this view was not correct. It is only the *well developed* crystals that are exceptional. A very large number of solids which do not appear to be crystalline at first sight are seen to be composed of tiny crystals when examined under the microscope. The list includes all rocks and all metals; only a small group of recalcitrant solids such as glass fail to show any trace of crystallinity.

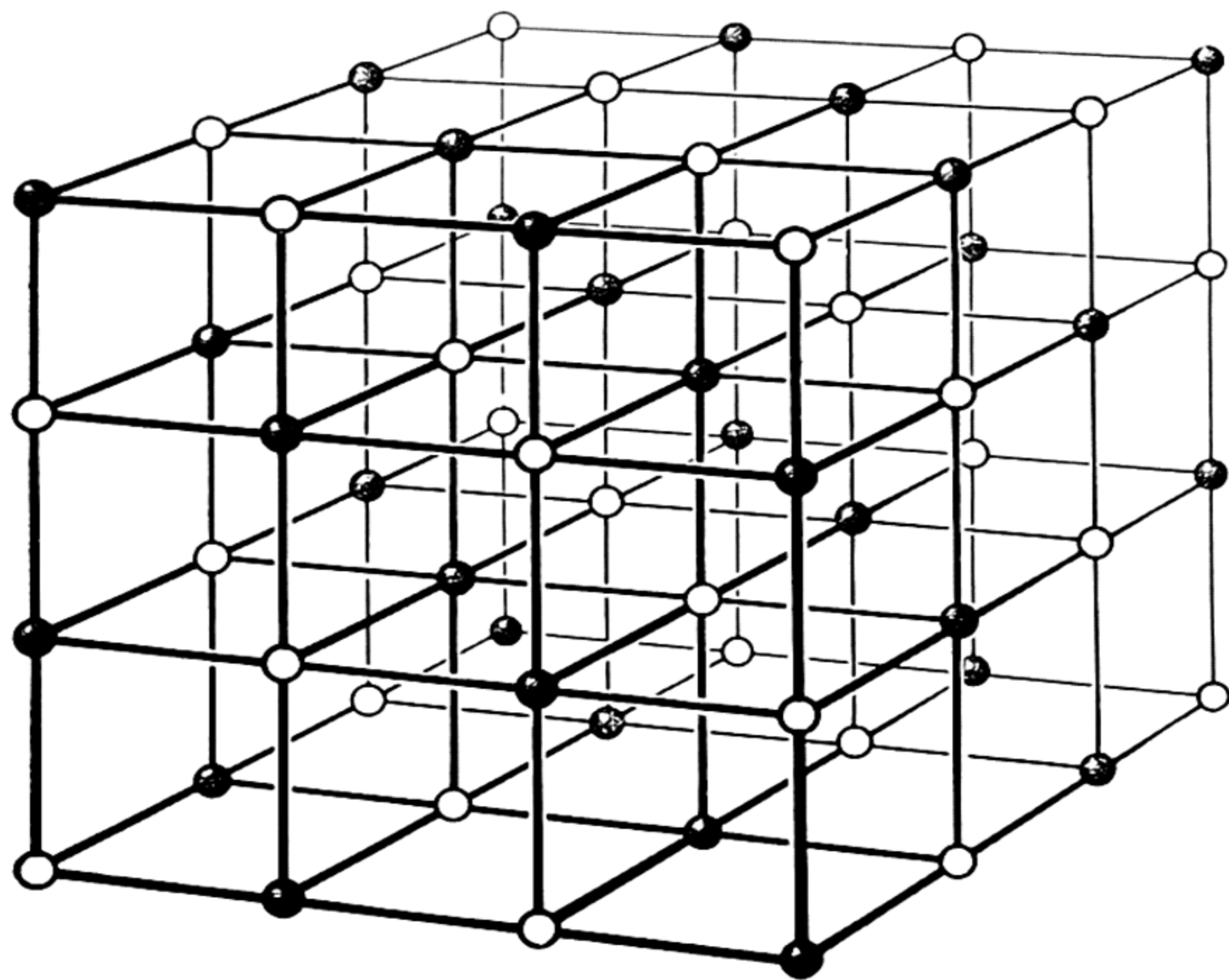
Crystal Structure

Crystals are homogeneous solids bounded by plane faces. Many of them are strikingly symmetrical in shape; for

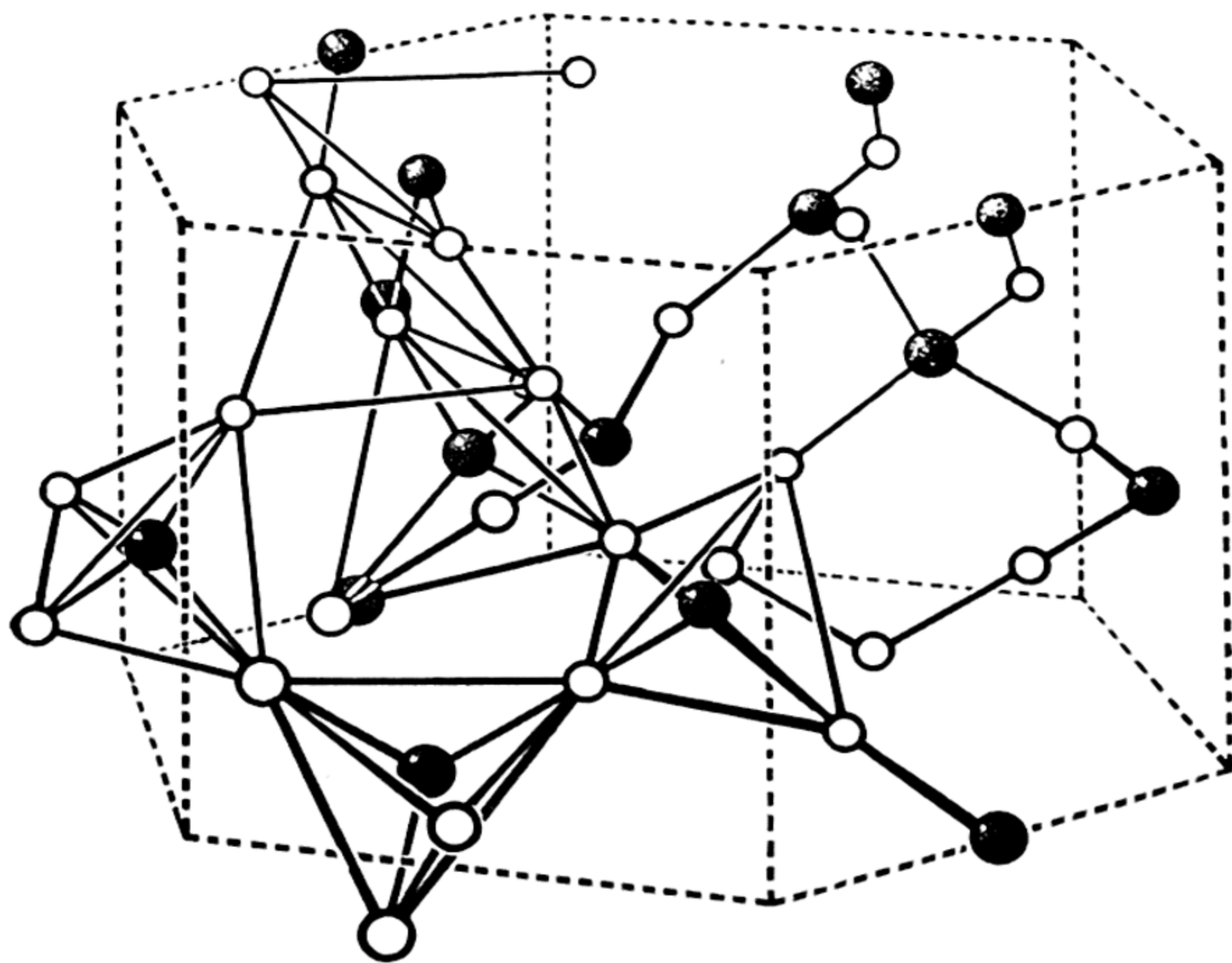


CRYSTAL owes its characteristic shape to the regular geometry of its atomic structure. This crystal is ethyl-

ene diamine tartrate (EDT), which is grown in solution and cut into sections to control electrical frequencies.



COMMON SALT or sodium chloride crystal is made up of cubic units. The bonds between sodium (*black*) and chlorine (*white*) are marked by lines.



QUARTZ or silicon dioxide has a basic hexagonal structure determined by the angles of the bonds between silicon (*black*) and oxygen (*white*).

example, the little cubes formed by rock salt and the hexagonal prisms of quartz. Even in some of those that are not fully symmetrical, the asymmetry is often due simply to the fact that certain planes have been shifted parallel to themselves. If the angles between the planes are taken as the basic indication of structure, the same structure always goes with the same chemical species. For this reason crystals acquired great value to the chemist as a means of identifying and preparing his chemical compounds. He had only to permit the substance to crystallize out of a solution or a melt; it then corresponded (with some exceptions) to a simple chemical formula. Thus there was added to the geometric simplicity of crystals their simple chemical behavior.

It was inconceivable that this combination of properties was not due to a symmetry in the internal constitution. A regular arrangement of molecules was postulated, and the consequences of that postulate were worked out by mathematicians with the tool of group theory. They proved in particular that any crystalline symmetry can be obtained by reproducing the same unit over and over through identical parallel displacements, like so many houses in a modern housing development. This unit, called the unit cell, is not necessarily identical with the individual molecule; indeed, we know today many cases in which there are several molecules per unit cell. The parallel displacements form a lattice, the so-called Bravais lattice. The final triumph of this viewpoint was achieved by the proof that an arrangement based on molecular symmetry could never exhibit the five-sided symmetry which is so common in living nature (*see drawings on page 419*).

When it became possible to measure the positions of atoms in crystals with X-rays, these crystallographic theories became one of the foundations of the new physics of the solid state. The discovery of this technique was made in 1912 by Max von Laue of the University of Munich and his associates Walter Friedrich and C. M. Paul Knipping. The experiment, one of the most important of all time, arose in the following way. Von Laue, a member of the laboratory of Wilhelm Röntgen, the discoverer of X-rays, was inclined to consider these rays as electromagnetic waves; that is, of the same nature as radio waves or visible light, but of much shorter wavelength. To prove that they were waves he had to construct or find gratings much finer than any grating ever made by man. At this point he happened to become acquainted with studies then being made of crystals, and he perceived that here was a grating made by nature which should serve his purpose. At his suggestion Friedrich and Knipping directed a beam of X-rays at a crystal, and looked for reflections

from the regular crystal planes which existed in the interior. After some unsuccessful trials these reflections were found.

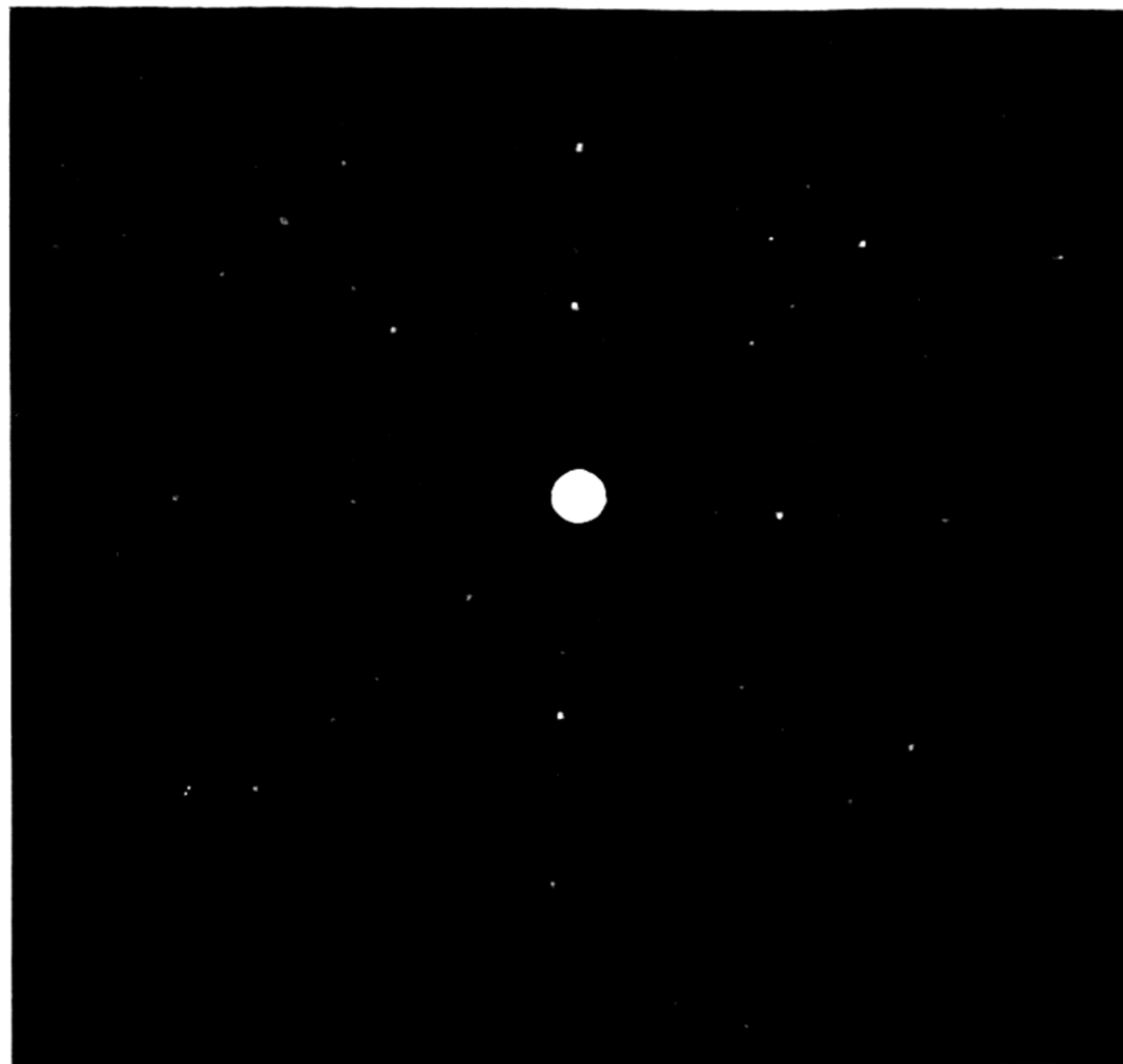
Within a few years the technique of diffracting X-rays by means of crystals was taken up by others, notably W. H. and W. L. Bragg, father and son, in England. Today it has become a standard tool for analyzing crystal structure, while its original purpose—determination of the wavelength of X-rays—has receded into the background. The simplest example, and the first to be analyzed with this tool, is the sodium chloride crystal, NaCl (*see drawing on the opposite page*). The partners Na and Cl occupy the corners of a cubic lattice; they alternate in this position, like the black and white squares of a checkerboard. Each Na is surrounded impartially by six Cl's at equal distance, and *vice versa*.

Thus crystallography provides a geometric analysis of the solid state which is unusual in its beauty and perfection. But it is not yet physics. Johann Kepler's laws of planetary motion, which had a similar beauty, were not physics but astronomy; Newton transformed them into physics by finding the law of force to which the planets were subject. In the same way physicists asked what forces made the atoms in crystals arrange themselves as they did, and what dynamic phenomena took place in crystals. They learned that the forces responsible for the formation of atoms, molecules and crystals are electrical, which placed solids and molecules on a similar footing.

The Bond

It is therefore appropriate to examine the chemical binding forces in molecules as a preliminary to the more complicated case of solids. The basic unit of these forces is the negatively charged electron. Even before the discovery of the electron by the Cambridge physicist J. J. Thomson in 1895 the existence of units of electrical charge had been detected in solutions. When a salt such as sodium chloride is dissolved in water, the molecule dissociates into sodium atoms, charged positive, and chlorine atoms, charged negative. These charged carriers, called ions, are identified by the passage of an electric current, in which a transport of charge is linked in a fixed way with deposition of matter—atoms of sodium on one electrode and atoms of chlorine on the other. Thomson's discovery that the charge of the electron was identical with the known unit of charge of ions suggested immediately that electrons were constituents of the atom and that ions were characterized by an excess or a defect of such particles.

At first the electrons in the atom were thought of as negative raisins in a positively-charged cake; then it was shown



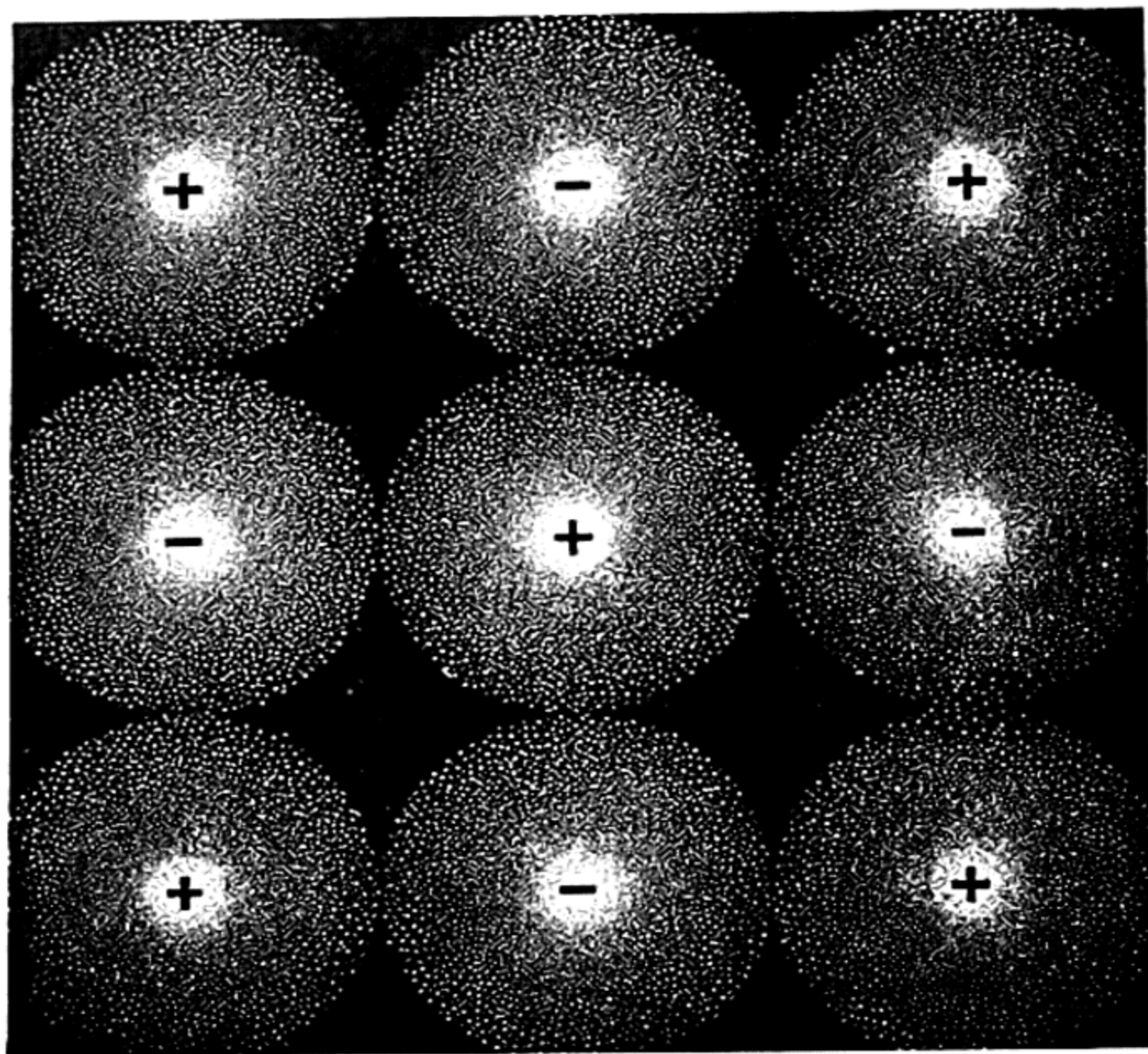
X-RAY DIFFRACTION PATTERN of a barium titanate crystal has a regular array of spots. The hole in the center is for mounting the negative.

that the positive charge, far from being a cake, is concentrated in an extremely small space within the atom where it forms the atomic nucleus, about which the electrons cruise very much like the planets around the sun. Next it was seen that the charge of this nucleus determines the chemical element. Thus the atom of copper consists of a nucleus bearing 29 positive charges surrounded by 29 negative electrons; from it an ion with two extra units of positive charge (Cu^{++}) is derived by the surrender of two electrons. The manner in which these electrons are bound to the nucleus was elucidated only after considerable difficulty. It was learned that the electrons were restricted to certain definite orbits, representing energy states, around the nucleus. A modification of mechanics, called quantum or wave mechanics, had to be evolved to pick out the possible orbits; between these possible states the electrons can make transitions, emitting or absorbing the energy difference in the form of light.

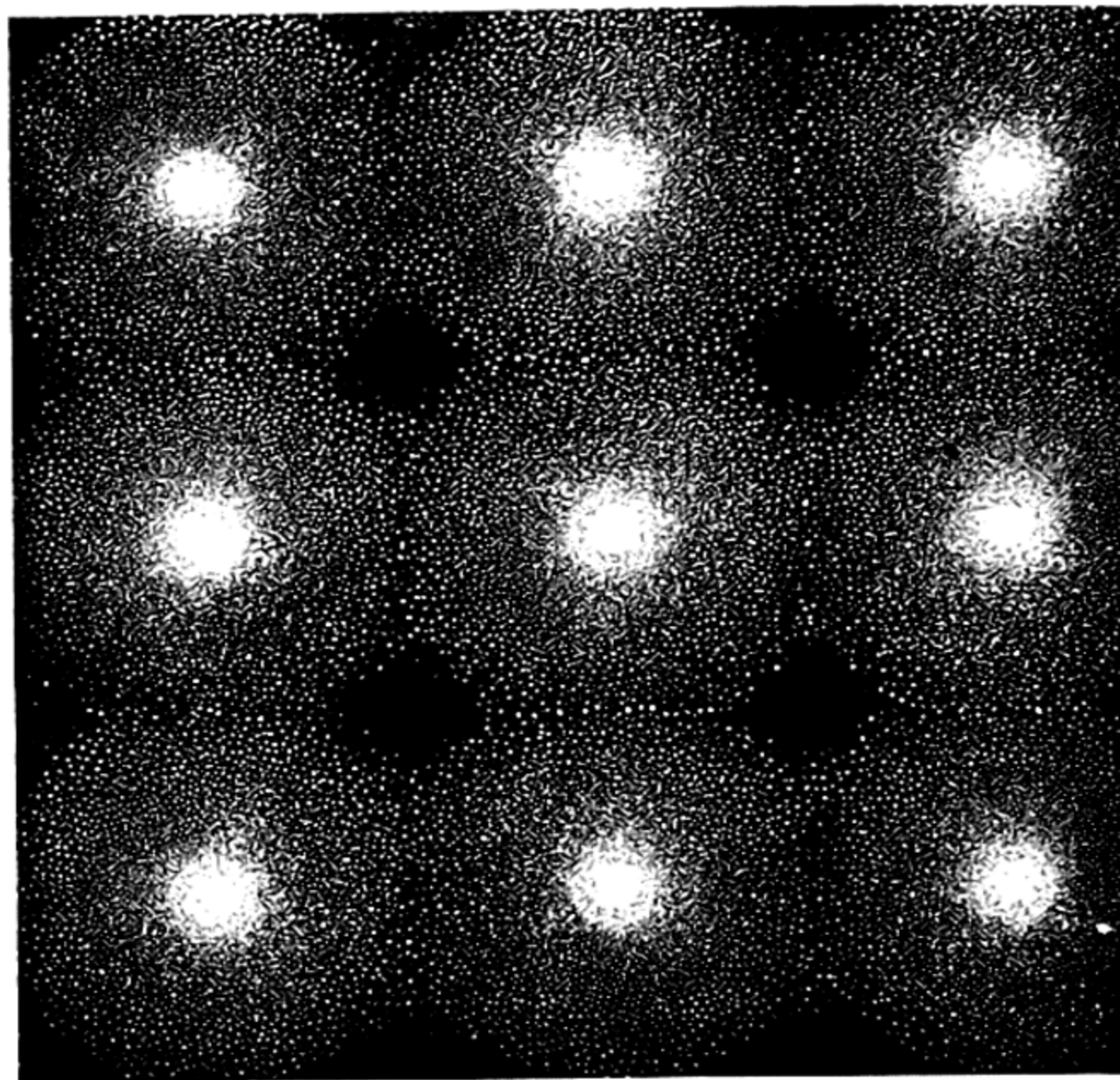
The picture was reasonably well completed by the discovery in 1927 that the electron spins on its own axis, much as the earth does, and that only two states of rotation are possible. With this discovery, it became possible to formulate the so-called Pauli exclusion principle, which states that two electrons cannot be in the same quantum state including spin. The Pauli principle gives the elec-

trons around the atom a shell structure, and so explains valency.

Thus the lightest atom, hydrogen, which consists of a nucleus bearing one positive charge and an electron, has the valency 1, corresponding to the fact that it can surrender its electron to an atom desiring one, as it does in hydrochloric acid (HCl). The next atom in order of weight, helium, has two charges and two electrons. This arrangement is so stable that the electrons do not usually leave the atom and helium does not form chemical compounds. The next atom, lithium, has three electrons, only two of which can enter the stable helium shell, while the third, called the valence electron, is easily detached. Consequently lithium has valence 1 and forms salts such as lithium chloride, in which lithium bears a positive charge; a solution of this salt in water will conduct electric current, with lithium coming out at the negative pole. Beryllium, the next element, has two valencies, and so on up the scale of elements. The chemically inert gas argon, with 10 electrons, fills the second shell. The element immediately preceding argon, fluorine, lacks one electron in the second shell and is an extremely reactive substance with one valence. The valence is of the electronegative type; that is, fluorine tends to grab an electron to complete its second shell. In consequence a salt such as lithium fluoride, in which the lithium



IONIC CRYSTAL is composed of atoms held together by opposite electric charge. The shading of each atom represents the density of the electron cloud about its nucleus.



COVALENT CRYSTAL is composed of atoms that share their outer electrons. The outer part of each electron cloud joins that of its neighbor.

atom passes one electron to the fluorine atom, is very stable, because both atoms are now surrounded by complete electron shells. After argon comes sodium, with 11 electrons, one of which is detached easily; this means that it will act very much like lithium. Thus the elements fall into groups of similar behavior according to the structure of the outermost shell. The group of atoms having just one electron outside a closed shell is called the alkali metals. This group occupies a key position in the theory of the solid state, as will be seen.

Crystal Bonds

Now let us return to the structure of crystals. The simple lattice of the sodium chloride crystal is composed of ions. Ionic crystals had revealed something of their electrical character as early as the 18th century. Jewelers, testing the durability of the colorful gem tourmaline by putting it in the fire, noticed that foreign particles collected on it, and they named it "the stone which attracts ashes." Careful study showed that the gem became electrified when it was heated, always in the same crystallographic direction, positive charges appearing on one crystal face and negative charges on its opposite. More than a century later French scientists followed up this discovery of a "pyroelectric" effect with demonstration of the "piezoelectric" effect—the electric polarization of certain crystals under pressure. These polarizations result from shifts in the equilibrium positions of the ions; the ions return to their original positions when the heat or the squeeze is re-

moved. The shift of positive ions is different from that of negative ones. Today piezoelectricity is sufficiently well understood to have made possible the design of artificial crystals which can replace natural quartz in some applications ["Crystals and Electricity," by Walter G. Cady; *SCIENTIFIC AMERICAN*, December, 1949].

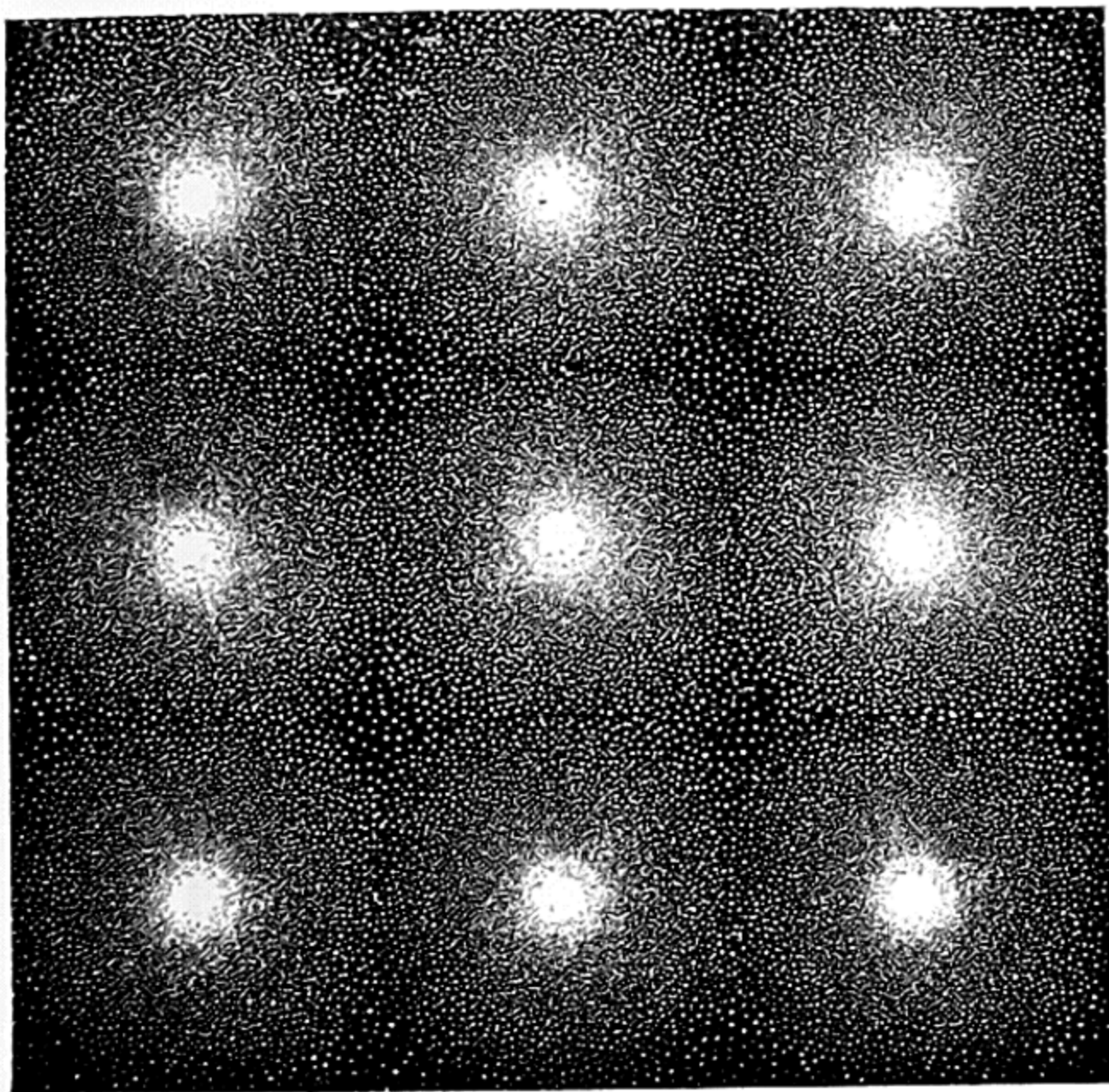
The ionic crystals, and sodium chloride in particular, were not only the first solids whose structure was analyzed, but also the first chemical compounds whose chemical binding energy was accounted for by physical principles. This feat was accomplished by the German scientists Fritz Haber and Max Born. Haber suggested that the chemical binding of sodium and chlorine in the rock salt crystal was accomplished simply by the electric attraction of Na^+ and Cl^- . If this hypothesis was true, the chemical binding energy could be computed in electrical terms, and Born was able to do so. His calculation was the first computation of a chemical quantity from physical premises. The importance of this step cannot be exaggerated. One can infer from it that very probably all chemical binding is electrical in the last analysis, and that it should be possible to implement this view by direct computation.

To explain how solids hang together, however, one missing link remains to be forged. The ionic concept does not account for the binding of atoms in all chemical compounds. In the molecules of hydrogen (H_2), oxygen (O_2) or chlorine (Cl_2), for example, neither of the two atoms which compose the molecule

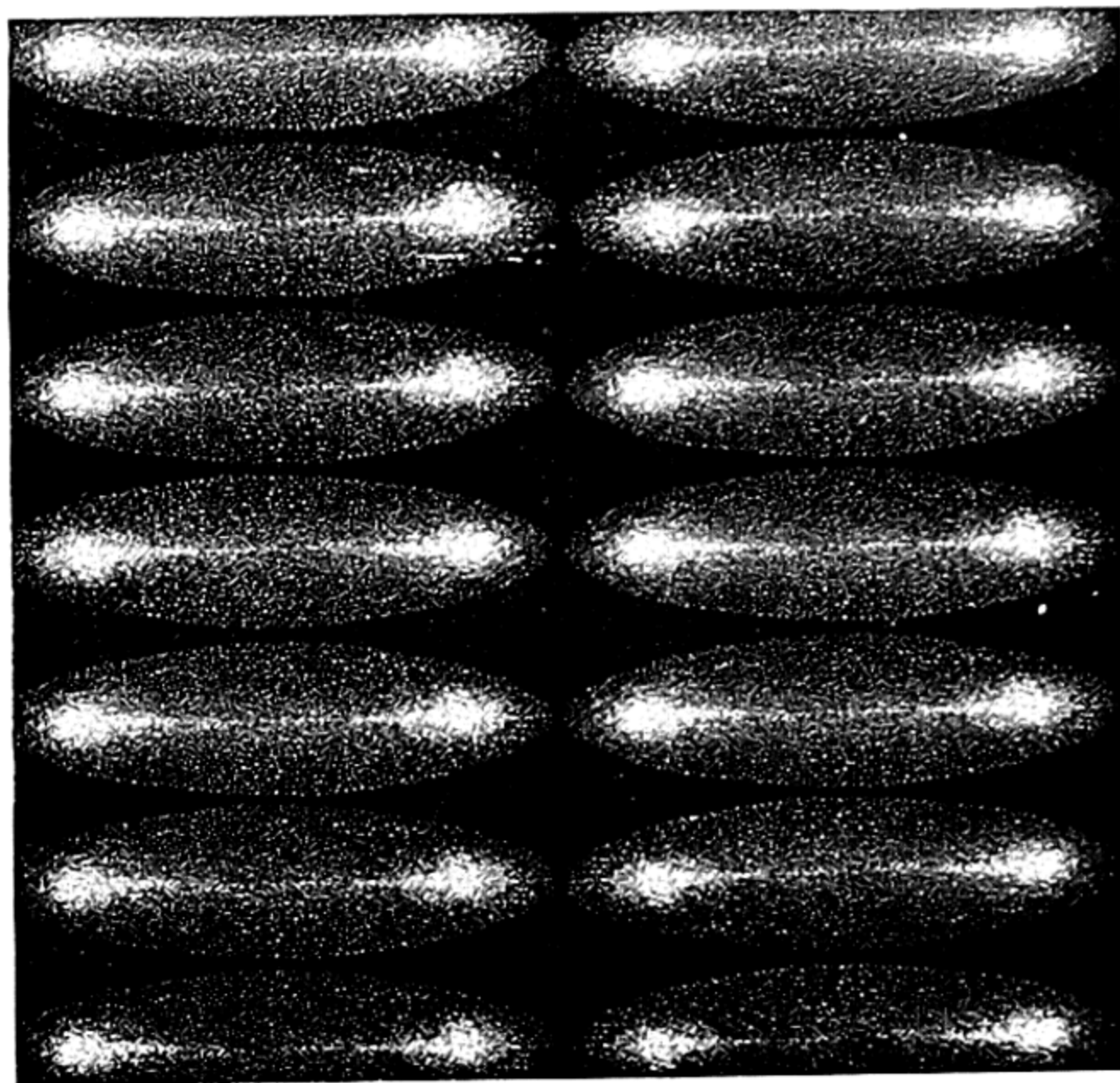
carries an effective charge. Similarly in many metallic crystals all the atoms are known to be in the same condition, and in consequence, we cannot possibly assume half of them to have positive charge and half negative. The chemical bond in such cases is called covalent. It is one of the successes of quantum mechanics that it can explain the covalent bond; because of its importance for the theory of the solid state as well as for chemistry we must examine that bond in some detail.

The Covalent Bond

Quantum mechanics, sometimes called wave mechanics, represents the final formulation of a schizophrenic viewpoint which has evolved in 20th century physics. Earlier scientists heatedly debated whether light was made up of particles or waves. This conflict finally was resolved by the assertion that it is both. There is a discrete unit of light called a quantum, but if we try to locate this particle, we must apply wave theory. The height of the wave at any point gives the probability of finding the light quantum at that spot. Wave mechanics is concerned with the second stage of this reasoning. If light waves are also particles, then particles are also waves. The wave pattern associated with a particle gives the probability of finding it at a given spot: on the wave crests the probability is high, in the troughs it is low. The wavelength of this "probability wave" is obtained from the formula devised by the French physicist Louis de Broglie; the wavelength equals Planck's constant divided by the mass



METALLIC CRYSTAL is composed of atoms whose outer electrons are so loosely held that they are free to move through the crystal lattice.



MOLECULAR CRYSTAL is made up of molecules. These are held together by weak forces due to the attraction of nuclei in one molecule for the electrons of another.

times the speed of the particle. It follows from this formula that a particle cannot be located with certainty within a space much smaller than one de Broglie wavelength. Therefore you cannot confine the particle without doing work, because in order to do so you have to reduce its wavelength. This reduction can be accomplished in de Broglie's formula in only one way: by increasing the speed of the particle. An increase in speed means an increase in energy, which has to be supplied from somewhere.

For electrons this feature of quantum mechanics is particularly important, because the mass of an electron is far less than the mass of an atom and hence the wavelength tends to be larger. The compression of an electron into a space of atomic dimensions therefore requires significant amounts of energy. If an electron is given a chance to spread itself over several atoms it will do so, and, between two possible states of matter, the one in which the electron has that chance will be the more stable, other things being equal. The question remains: What limits this process? Why do not all electrons spread out their waves indefinitely? The answer is Pauli's exclusion principle. Each possible wave pattern is a quantum state in the sense described for atoms, and according to the exclusion principle only two electrons (for the two directions of spin) can spread themselves out in the lowest energy state, or wave pattern. The third and fourth electron must go into the next higher state, and so on. The tendency to form covalent bonds can be de-

scribed as the tendency of the electrons to extend their wave patterns, and thus to reduce their energy of motion, as far as this is compatible with the Pauli exclusion principle.

Take as an example the simplest molecule, H_2 , formed by two atoms of hydrogen. The system contains only four constituents: two positively charged nuclei and two electrons. When the two atoms are far removed from each other, the electric forces confine each electron to its own nucleus. If the two atoms are brought closer, the wave corresponding to each electron can be spread over the two nuclei, lowering their energy. The Pauli principle can be satisfied by giving the electrons opposite spin. In consequence the energy of the system as a whole is lowered, and the two atoms cannot be separated again without furnishing a certain amount of energy; this is the chemical binding energy of the molecule H_2 .

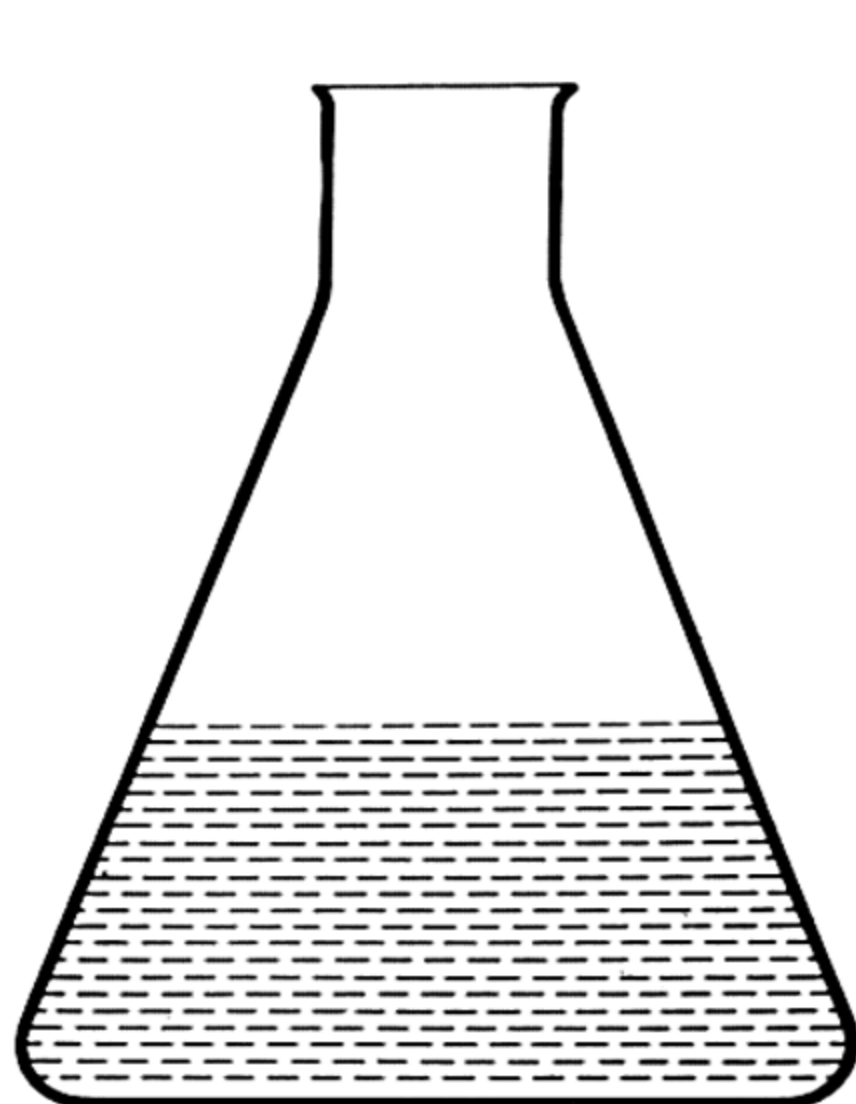
Conductors and Nonconductors

The covalent two-electron bond, the bond achieved by two electrons of opposite spin sharing their wave function, is the most frequent bond of chemistry. It is much more usual than the ionic bond in which an electron is transferred from one atom to another. Its widest application is in organic chemistry. On the other hand, very few crystalline solids are built exclusively on this principle. The classical examples are diamond, germanium and gray tin, a nonmetallic substance. In these crystals the number of nearest neighbors of any given atom just equals its valence. This situation is

most perfectly exemplified in the structure of diamond, made up of four-valent carbon. But in general covalent bonding enters into the theory of crystals in a secondary way, in conjunction with other forms of binding.

Implicit in all this is the fact that covalent as well as molecular and ionic crystals are essentially nonconductors of electricity. To be sure, the charged ions in ionic crystals can conduct electricity, but generally their conductivity is small. Electric conduction demands mobile electrons. In these three types of crystals the electrons are all locked into certain quantum states. Metals represent a fourth type of crystal, in which the number of quantum states must be greater than the number of electrons available to fill them, giving the electrons freedom to switch and to move around. For example, in metallic lithium each atom is surrounded symmetrically by eight other atoms. The number of valence electrons per atom is one. If there were no electron spin, one quantum state on each atom would be filled. However, because of spin the number of available states is double the number of electrons, thus giving them a possibility of motion. Because of the electronic charge this motion can be observed as electric current.

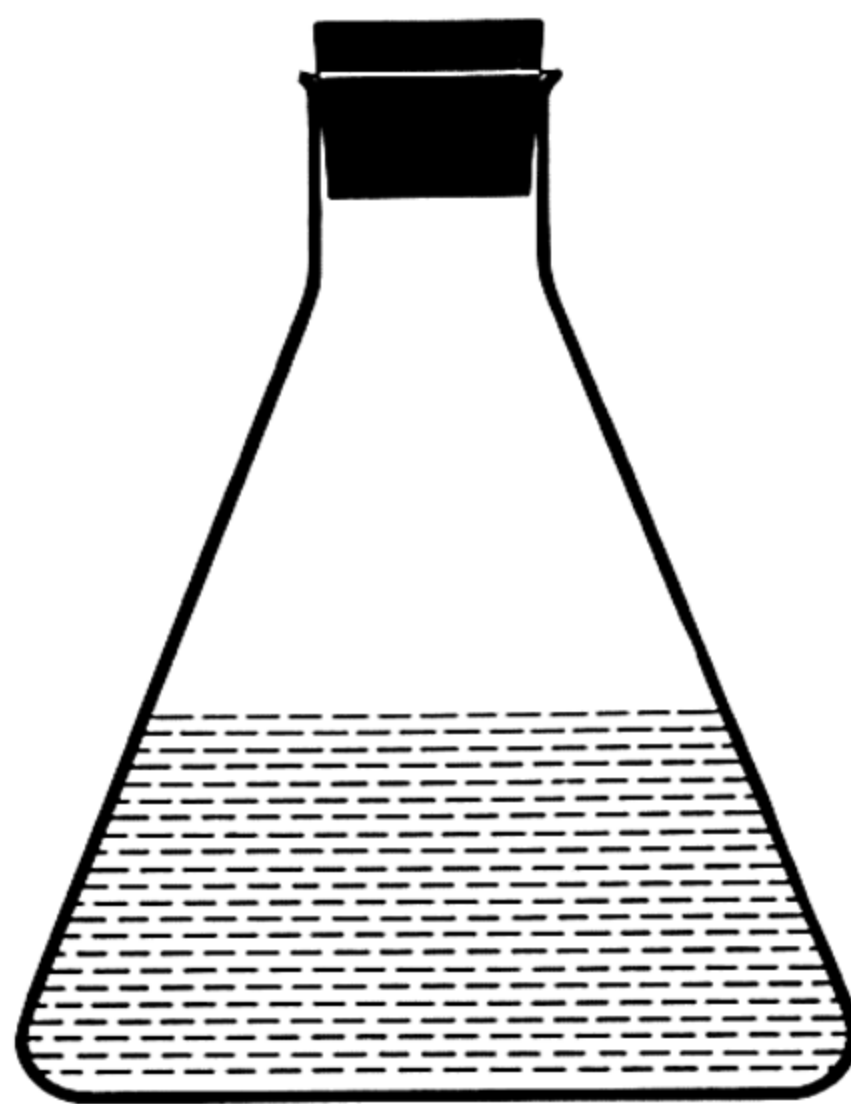
In 1900, five years after the discovery of electrons, the German physicist Paul Karl Ludwig Drude suggested that they were the agents which conduct electricity in metals, and he constructed a theory in which he assumed the electrons in the metal to be free, like the molecules in a gas. His theory assumed that the flow of current in a metal depended



FREE ELECTRONS are roughly analogous to water in half-filled vessel.

on two factors: the applied electric force and the resistance offered by the numerous collisions between the electrons and the atoms in the metal. Drude's formula was not immediately verifiable because it contained an unknown—the mean free path of the electrons between collisions. However, his theory could be used to calculate the conductivity of heat as well as that of electricity. The ratio between the two conductivities could be computed theoretically and compared with experiment. The result was that theory and experiment checked each other very well.

Yet Drude's theory raised more questions than it answered. One troubling question had to do with the matter of specific heat. If the electrons form a gas, they must obey the laws that apply to gases. The specific heat of a gas is easily computed from theory, because it tells exactly what energy is required to impart to the gas molecules a given speed.

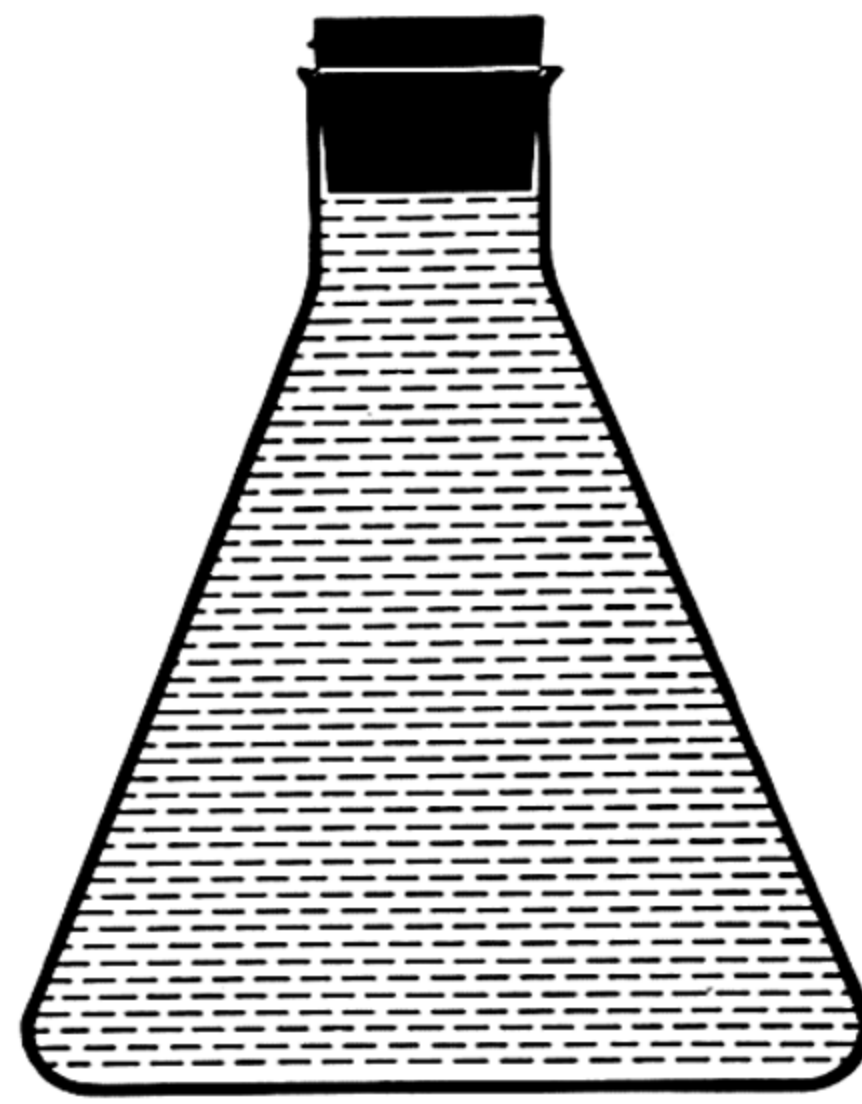


ELECTRONS IN SODIUM behave like water in vessel that is stoppered.

These same considerations, applied to a gas of free electrons, predicted that the specific heat of a metal should be larger by a substantial amount than the specific heat of an insulator. But no such difference was found. This contradiction, together with the fact that no criterion was forthcoming to tell when electrons are free and when they are not, stopped further progress. The Drude theory remained a theoretical fragment of doubtful validity which sometimes gave right answers and sometimes did not.

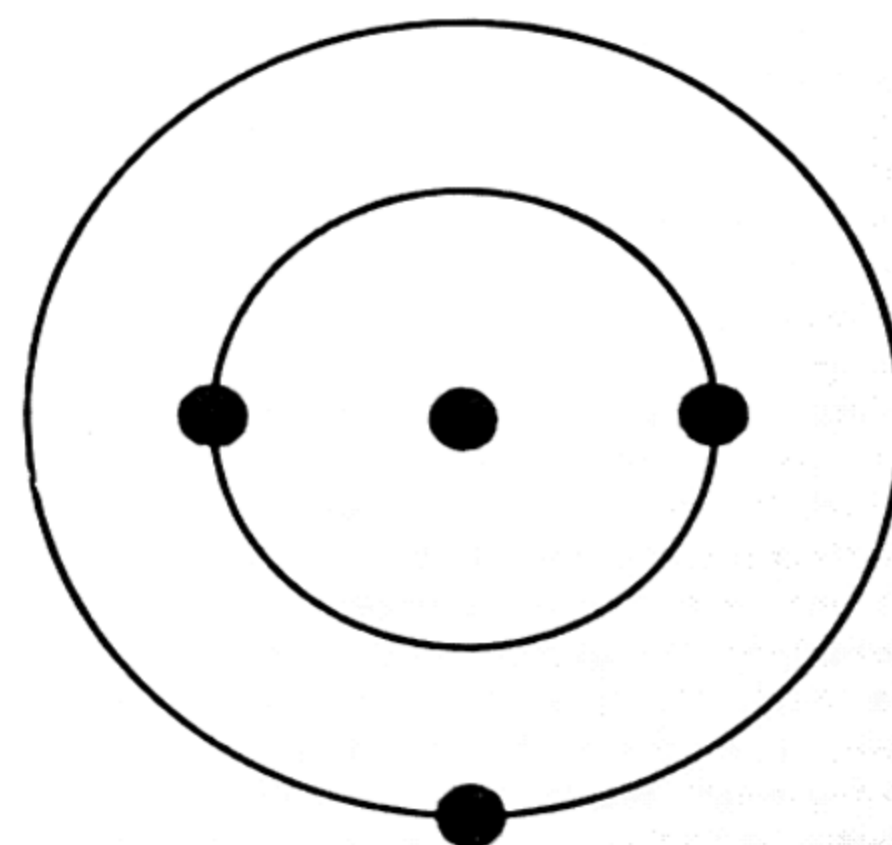
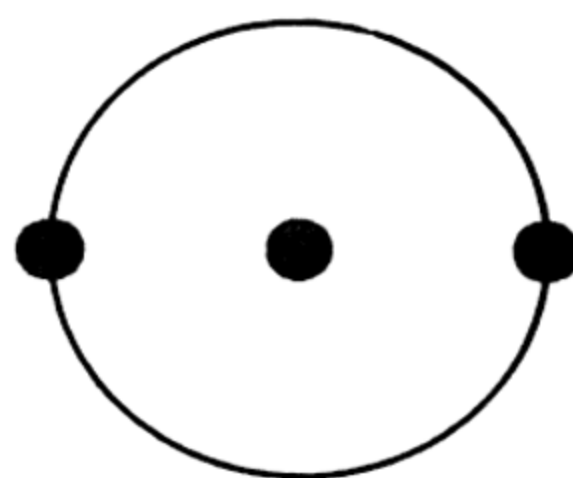
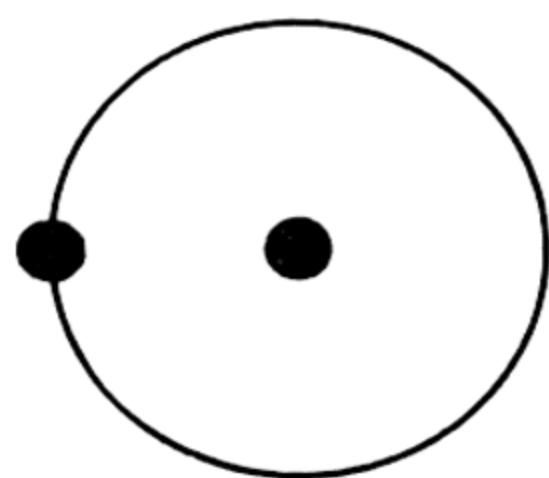
The Electron Gas

The development of quantum mechanics and the enunciation of the exclusion principle resolved the apparent contradiction within the Drude theory. The solution was found in 1928 by Arnold Sommerfeld of the University of Munich. Applying quantum mechanics to the hypothetical electron gas of Drude, he showed that electrons, even



ELECTRONS IN DIAMOND behave like water in filled, stoppered vessel.

at room temperature, are in a condition where all low-lying quantum states are tightly occupied. A gas in such a condition is called a degenerate gas. If we try to heat our degenerate electron gas, we find that most electrons are incapable of accepting energy because all neighboring states are occupied. Only a small number of electrons of highest energy can accept heat in the usual way. In many ways the degenerate electron gas inside a metal can be compared to a container filled with water in which the energy is simulated by the height, and Pauli's exclusion principle by the impossibility of two drops being in the same place. What corresponds to the surface level is the so-called Fermi level for electrons. Below the Fermi level all states are filled; above they are all empty. If the container is agitated slightly, only the surface drops can jump up in the air; the other drops are hindered by the ones above them.



ELECTRON SHELLS account for the repetition of chemical properties in the table of the elements. Hydrogen (*left*), has one elec-

tron traveling in one shell. Helium (*middle*), has two electrons which fill shell. Lithium (*right*), has third electron which travels in second shell.

A final clarification of the matter was worked out by Felix Bloch, Eugene P. Wigner and Frederick Seitz, now respectively at Stanford and Princeton Universities and University of Illinois. They studied chiefly sodium. Think of the metal as arising from a gradual pushing together of independent atoms. As the atoms approach one another, they arrive at a distance at which the valence electrons can jump from atom to atom. De Broglie's relation then enters into play: each electron tries to reduce its energy by spreading its wave function uniformly over the entire metal. But according to the Pauli exclusion principle only two electrons in the entire metal can do so. The other electrons have to accept states with shorter wavelengths and a correspondingly smaller binding energy. Thus the electrons are piled on top of each other just as in the degenerate electron gas. The totality of states available to these electrons occupies a band of energy, and the filling occurs to a certain level. In the case of sodium there are exactly twice as many possible states as electrons, because there is one electron per atom outside a closed shell and two possible directions of spin. The metallic binding is a consequence of this sharing of electrons, and the conductivity a consequence of the incomplete filling of the available energy levels.

Returning to the water-in-the-container picture, we would say that the electrons in sodium are analogous to a bottle half filled with water, whereas the electrons in a valence crystal like diamond correspond to a full bottle. Finally the picture of free electrons is analogous to water in an open container. So long as the container, whether open or closed, is only partly filled, a slight tilting of the bottle—which simulates the effect of an applied static electric field—brings about a rearrangement of the electrons, emptying certain states and filling others. The filled bottle, on the other hand, shows no response. Thus a close analogy exists between free electrons and electrons in a partially filled band. It is clear that we are dealing with an analogy only; there is no guarantee that it is a perfect one. Indeed, we know it is not quite perfect, for to make it work one sometimes has to ascribe to electrons in a metal an artificial mass which is not their true value. But apart from this the analogy works rather better than one would have a right to expect.

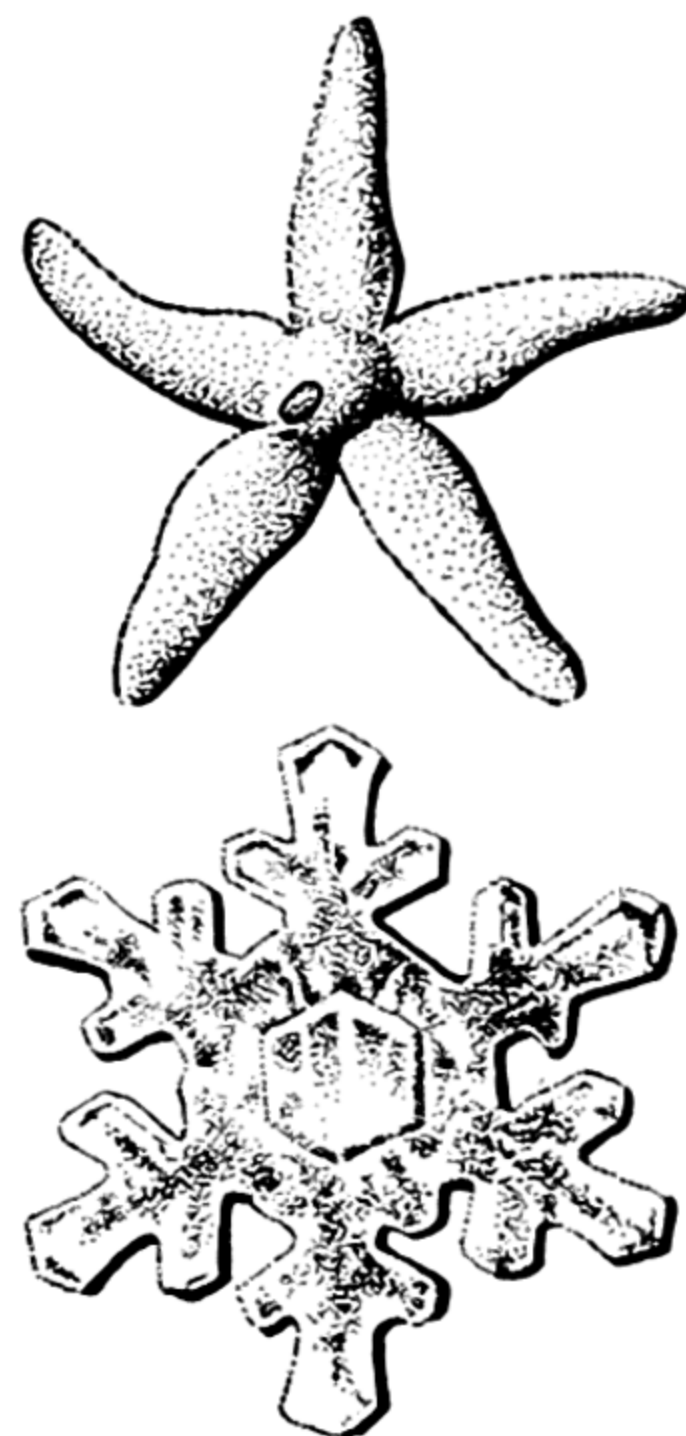
The band into which the quantum state of the valence electron in sodium spreads out is only an instance of a more general feature applicable to all quantum states of an atom when they are brought together. In a solid each atomic state becomes a band, with the lower ones generally narrower than the upper ones. Somewhere within this system lies the Fermi level. Below this level all en-

ergy states are filled; above it all are empty. If this level falls within a band, the body is a conductor; if it falls in a gap between bands, it is an insulator. We may say that the solid is similar to a bottle consisting of separated sections connected by narrow tubes. The water level in this bottle may lie within a section, in which case a slight tilting of the bottle can rearrange the regions occupied by the water near the surface (conductor), or it may lie within one of the narrow tubes, in which case tilting will have no effect unless it is very strong (insulator).

The division of solids into conductors and insulators has become complicated in recent years by the observation that all the so-called insulators conduct electricity to some extent. Some of them, called semiconductors, have enough conductivity to be technically interesting. Their conductivity is electronic, but in detail it is quite different from the metallic kind. Metallic conductivity increases as the body is cooled, because of the absence of impeding thermal agitation. Semiconductors, on the other hand, lose their conductivity at low temperature and gain upon heating. The explanation, given by the English physicist Alan H. Wilson, falls into the water-bottle picture developed here. If the tilting or the agitation in the bottle is violent enough, water may be brought up into the empty part, in compensation for which a bubble will appear in the full bulge. In other words, an electron appears in the empty band and a "hole" on the top of the full band. Such pairs, of electrons and holes are the carriers of electricity in a semiconductor. Since their number will increase rapidly with temperature, so, therefore, will the electrical conductivity of the material.

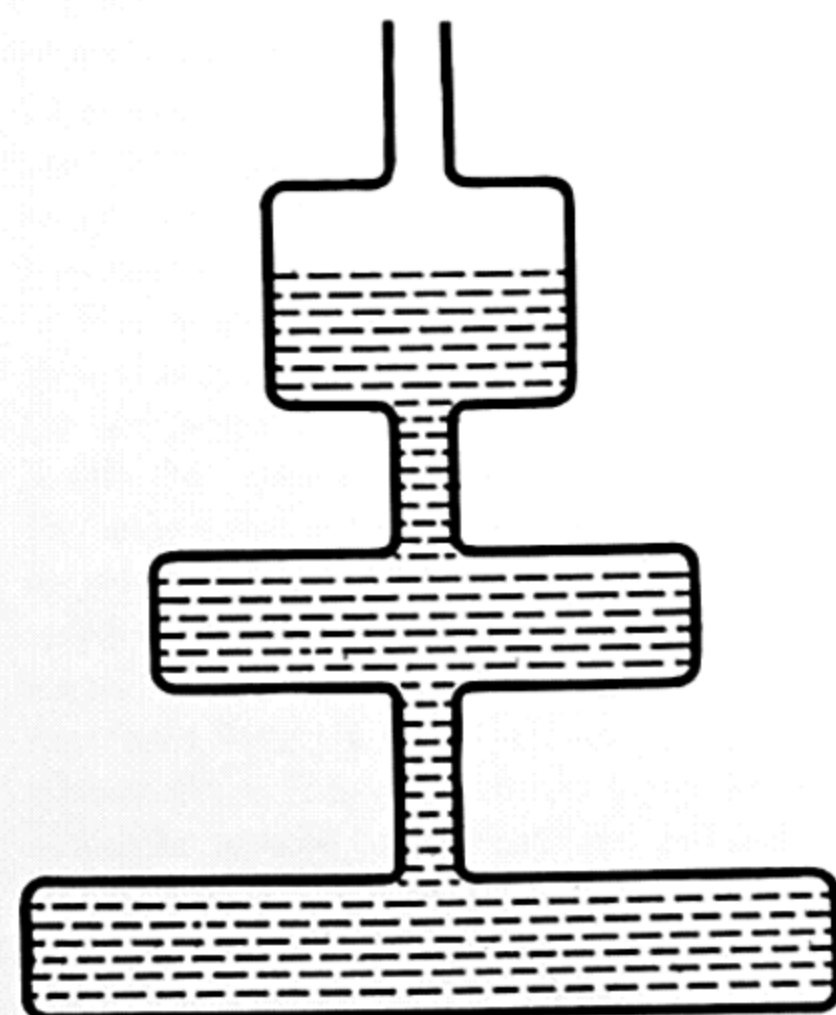
Assessing this conductivity one has a very easy time with the "excited" electrons. The Drude theory will apply to them in its original form, because they form a gas which this time is not degenerate; that is, the quantum states adjoining a given one are usually empty. One must not forget, however, the "hole" in the full band. The analogy with a bubble in a full water bottle is a particularly happy one in this connection. The bubble is able to move just as freely as the water in a way we can intuitively understand; it acts as if gravity were directed upward instead of down. In exactly the same sense a "hole" in a full band of negative electrons acts as if it had a positive charge.

Like all sciences that arise late, the physics of solids has the disadvantage of running behind technology. People have used and must use solids all the time, and in doing so, must acquire some sort of technological control over their properties. In such a situation the practical man is apt to despise the dreamers who try to "understand" fea-



FIVE-FOLD SYMMETRY of biological forms is never found in crystals. At top is a five-armed starfish; at bottom a six-sided 'snow crystal.

tures which he takes for granted. It looks, however, as if the time for this kind of attitude is about to run out. Theoretical analysis is beginning to produce solids with properties beyond the "common sense" experience of the practical man. The transistor is only one of the spectacular results that are destined to flow from this painstaking study of what has long seemed a dense and commonplace subject.



SECTIONS IN BOTTLE are analogous to energy levels in a crystal.

The Author

GREGORY H. WANNIER is a theoretical physicist at Bell Telephone Laboratories. He was born in Basel, Switzerland, in 1911. He received his undergraduate education at the University of Louvain in Belgium and at the University of Cambridge in England and took his doctorate in mathematical physics, home again, at the University of Basel. He came to the U. S. in 1936 as an American Exchange Fellow at Princeton University and, after several academic appointments, went into industrial research

with the Socony Vacuum Oil Company in 1946. At Bell Laboratories since 1949, he has been a member of the brilliant group of investigators who opened up the present fruitful period of study in the field of solid state.

Bibliography

CRYSTALS AND X-RAYS. Kathleen Lonsdale. G. Bell and Sons, Ltd., 1948.
THE MODERN THEORY OF SOLIDS. Frederick Seitz. McGraw-Hill Book Company, Inc., 1940.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound **SCIENTIFIC AMERICAN** Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

THE RADIO GALAXY

by Gart Westerhout

From somewhere near the center of our galaxy vast billows of hydrogen surge outward, emitting radio waves that indicate the galaxy's spiral structure.

During the past decade radio astronomers have been charting the sky on a new kind of map. At first glance this map looks like a contour map of a mountain range. The celestial contours, however, trace out the lines on which the sky shines with equal brightness in the wavelengths of the invisible radio spectrum. The outstanding feature of the topography of the radio sky, drawn in long sweeping contour lines, is the band of the Milky Way, which appears brighter to the radio telescope than it does to our eyes. Elsewhere on the map individual radio sources stand out as peaks of brightness, marked by more closely bunched concentric contours.

From study of such maps, made on several different radio wavelengths, astronomers are now deducing important features of the structure of our galaxy, hitherto invisible or obscured from view. The sky itself is a kind of map: a spherical projection at trigonometric "infinity" of the unbounded reaches of space that lie beyond. By various ingenious methods astronomers using light-gathering telescopes have translated this map into a three-dimensional picture of our galaxy. They have shown the galaxy to be a thin disk of stars with spiral arms. The center of the galaxy is located in the bright region of the Milky Way in the constellation Sagittarius. Over large regions of the sky, however, curtains of dust and gas hanging in space have dimmed or entirely blocked our view. These clouds are so thick and extensive that we can see only about a 20th of the galaxy with light-gathering telescopes. But interstellar clouds do not hamper radio telescopes any more than clouds in the terrestrial atmosphere interfere with radio transmission. These modern instruments open up the whole galaxy to view and permit us to look into its very center.

Of course we cannot "see" the heavens with a radio telescope. All we can see in a literal sense is the wagging needle of a voltmeter or the wiggling trace of a recording pen. But in the map of the radio sky constructed from these observations we are beginning to discern the fine structure of the galaxy. Measurements of the intensity and the Doppler shift of the spectral line emitted at the 21-centimeter wavelength by hydrogen in interstellar space have resolved the curtains of dust and gas into the spiral arms of the galaxy [*see illustration on page 426*]. Measurement of the Doppler shift in this radiation has also revealed the rotational motions in a large part of the galaxy. It was already known that the sun and the stars in its neighborhood rotate around the galactic center once in about 200 million years. Similar studies of the 21-centimeter radiation from other galaxies, such as the Great Nebula in Andromeda and Messier 33, have established the density of hydrogen and the rotational motions in these systems. Radio emissions from the center of our galaxy are now completing the picture of the spiral structure and promise to give us new insight into the evolutionary process that accounts for it. They indicate the presence of a large mass of ionized hydrogen gas at the center and show that great waves of hydrogen are rippling outward from the central region in the plane of the galaxy. At the outer edge of these waves, where their outward rush has lost its impetus, new stars seem to be forming. Their newborn energy ionizes the gas in a brilliant ring of radio emission encircling the center. Based as it is upon the first penetration of the galactic nucleus, this picture is still speculative, but it suggests the major outlines of the evolution of our changing galaxy.

Much of the early work of charting

the radio sky was done in the Netherlands. In fact, the first short-wave map of that part of the Milky Way which is visible in the Northern Hemisphere was made with the Leiden Observatory's 82-foot radio telescope located at Dwingeloo. When it first came into operation in 1956, this instrument was the largest of its kind. Now several radio telescopes in Europe and the U. S. equal it in size, and the reflector of the giant telescope at Jodrell Bank in England is more than three times larger.

The mounting of our telescope at Dwingeloo represents a departure from astronomical tradition. It is of "altazimuth" design, that is, the entire mounting rotates on a circular track around a vertical axis and the reflector pivots around a horizontal axis. The parabolic reflector is surfaced with a fine wire mesh and focuses the incoming radio waves on a small antenna mounted atop a 40-foot mast. From the antenna the signal travels to a 700-tube receiver and amplifier circuit designed to monitor the 21-centimeter line of un-ionized hydrogen, but tunable to the nearby wavelengths on which radiation from ionized hydrogen can be received.

The length of radio waves, millions of times longer than those of light, accounts for the large dimensions of a radio telescope. In general, the longer the wave, the larger the reflector needed to resolve its source. Our 82-foot reflector has a resolving power of .57 degree at a wavelength of 21 centimeters. This means that two sources of radiation closer together than .5 degree of arc are seen as a single source, and that an infinitely small source appears as a blur one degree across—twice the diameter of the moon as seen from the earth. As in the case of optical telescopes, however, a large reflecting surface confers large "light" gathering power to make the

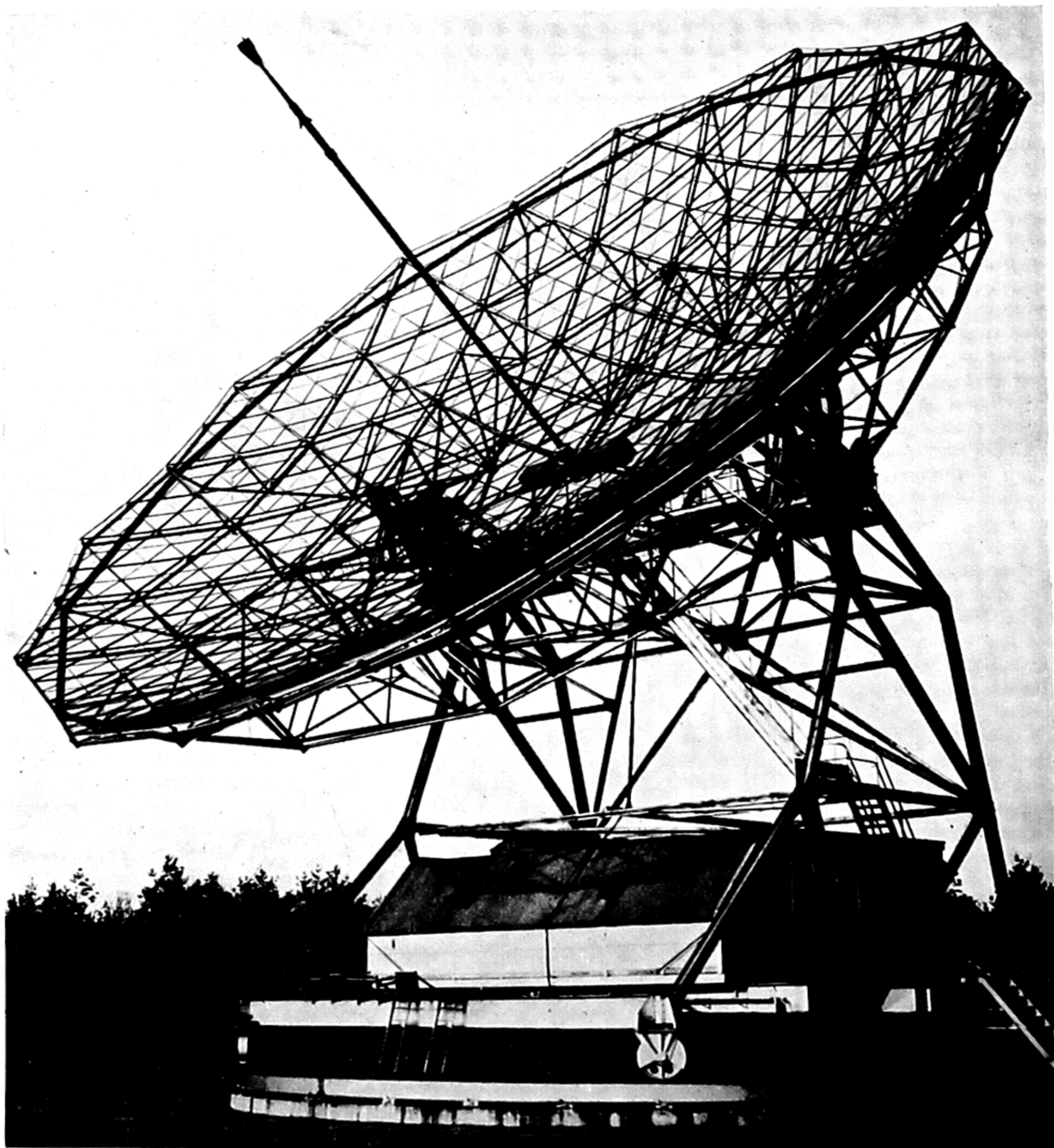
most of the tiny amounts of energy that reach us from any given point in the sky or from a discrete source in distant space. The telescope is sensitive to extremely faint radio sources and can locate them very accurately, although its accuracy compared to an optical telescope is heavily compromised by low resolving power.

In surveying the sky the telescope usually scans along the path of a star but eight times faster, starting behind

the star, then catching up with it and passing it, sweeping a line 30 to 40 degrees across the sky. After each such scan, the reflector swings a fraction of a degree to one side and repeats its sweep. The telescope thus sweeps across a section of the sky in much the same way as an electron beam sweeps across a television screen, but months of slow scanning with the radio telescope are required to produce a single sky map. The sky signal is amplified in the telescope

receiver and fed to a recorder, where it drives a pen tracing a continuous line on a moving strip of paper.

The effect of limited resolving power must be borne in mind in reading the record made by a typical scan with the telescope. The line drawn on the recorder corresponds to a strip of sky more than one degree wide, and it reflects only roughly the variations in signal strength that occur within a degree along its length. Thus it does not show stars or



RADIO TELESCOPE that helped to determine the structure of our galaxy is located at Dwingeloo in the Netherlands. Its antenna and the first stages of its radio receiver are mounted atop the mast that juts from the face of the 82-foot reflector. The telescope framework

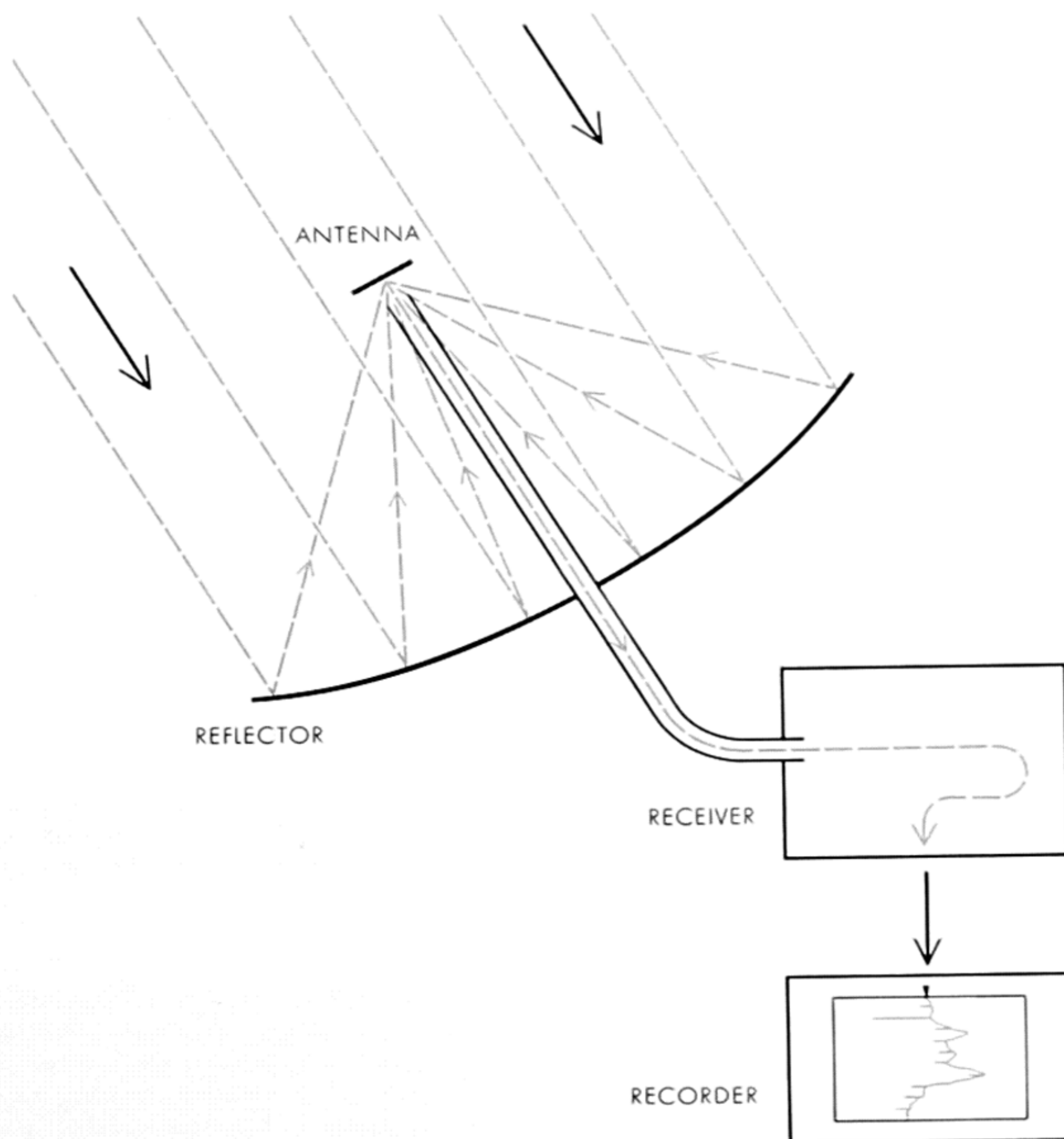
revolves on a horizontal track, while the reflector tilts up and down. The main part of the radio receiver is housed in the laboratory built into the supporting framework of the reflector; the laboratory building rotates slowly as the telescope reflector scans the heavens.

nebulae as point sources. The strong radio source Cygnus A, for example, shows up on a radio map as a blur with a diameter of more than a degree; it has been identified as a pair of colliding galaxies with photographic dimensions smaller than .01 degree. This should warn us that the many other radio sources which have an apparent diameter the same as that of Cygnus A are also of much smaller size. In order to smooth out errors a map such as the one on the next two pages is built up from hundreds of scans—two for every one-third degree.

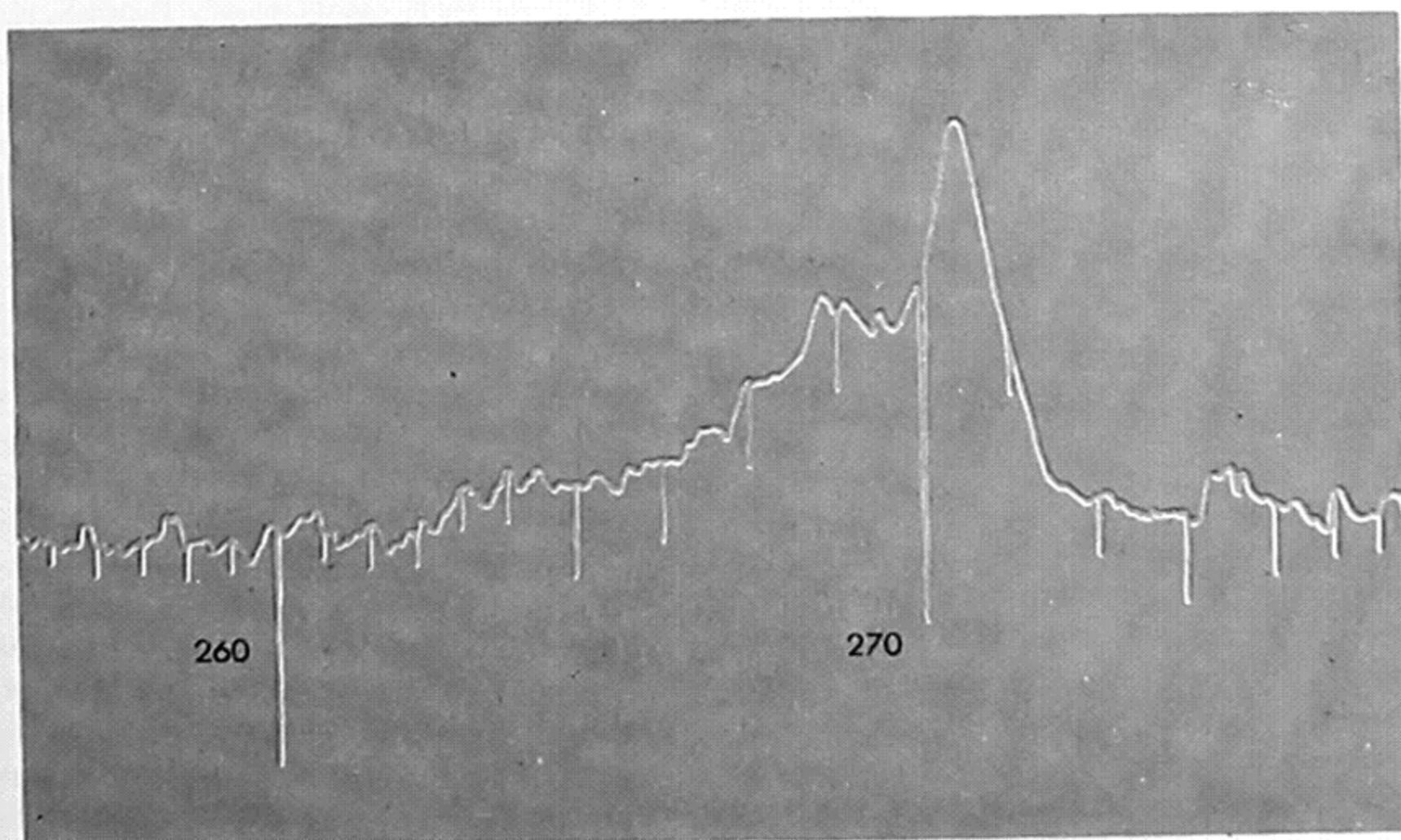
Radio maps made at different wavelengths differ significantly in what they show. On both long- and short-wave maps the Milky Way is the most striking region. A long-wave map (at, say, three meters and above) shows the Milky Way as a broad band across the sky, brightest in the region of Sagittarius. Bright radio peaks, the so-called discrete sources, cover the entire sky; so far some 2,000 of them have been located. While Cygnus A has been identified as an extragalactic source, the two other strong sources—the Crab Nebula and Cassiopeia A—have been identified as the remnants of stars that exploded within the galaxy. Are the other long-wave sources also exploding stars or colliding galaxies? No one seems to know. So far only a few of them have been identified with visible objects, although a considerable amount of detailed work remains to be done along this line.

On a short-wave map (25 centimeters and below) the Milky Way appears as a narrow ribbon, with the same ascendant brightness toward Sagittarius, but fading almost to invisibility near Orion. The discrete sources on this map are concentrated on and along the brightness ridge of the Milky Way; most of them do not coincide with the sources that appear on the long-wave map.

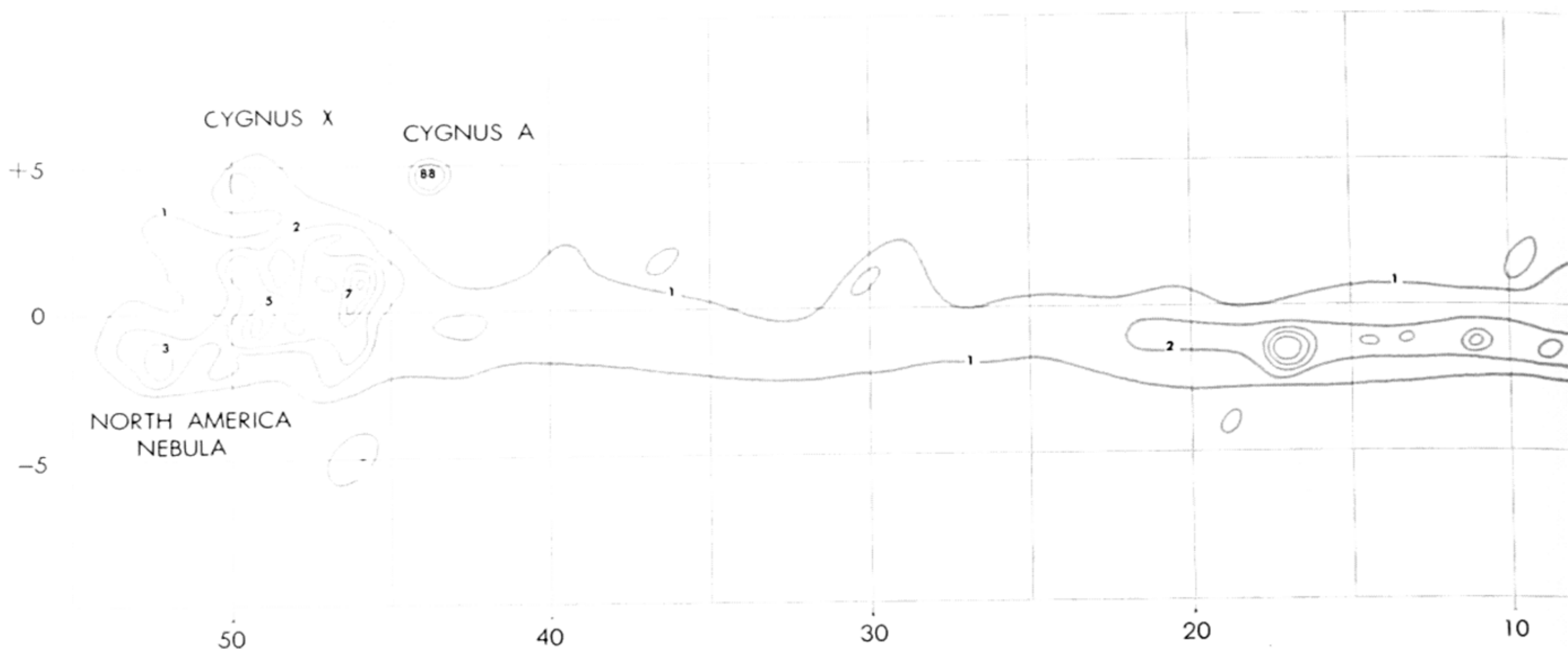
It is clear, therefore, that a complete radio atlas of the sky will require the making of many maps at many wavelengths. Radio astronomers all over the world are collaborating on a concerted effort to this end. At long wavelengths (3.5 to 15 meters) Australian observers have obtained remarkable results; British groups are working at wavelengths between 1 and 10 meters; an Ohio State University team has published an excellent 1.2-meter map [see "The Radio Sky," by John D. Kraus; *SCIENTIFIC AMERICAN*, July, 1956]. Our work at Dwingeloo continues in the wavelengths to which our instrument is adapted (21 to 75 centimeters), and the very short wavelengths (10 centimeters and be-



BLOCK DIAGRAM OF TELESCOPE shows how signal is picked up by the parabolic reflector and focused on the antenna. The signal is then amplified in a radio receiver and fed to a recorder, where a pen traces fluctuations in signal strength on a moving strip of paper.



TELESCOPE TRACINGS reflect variations in signal intensity along the telescope's path across the sky. This tracing is a cross section through the Milky Way; big hump is the nebula Messier 8, a bright radio source. Numbered dashes show degrees of right ascension.



RADIO MAP OF THE MILKY WAY was made by assembling a series of telescope tracings such as the one shown on the chart at the bottom of the preceding page. The contour lines on this map

represent lines of equal radio brightness; very strong radio sources appear as a series of closely spaced concentric contours. The Milky Way appears as a narrow band across the sky, dotted with bright

low) have been pioneered by workers at the Naval Research Laboratory in Washington, D.C.

Why the difference between radio maps of the sky? The answer is that radio signals that produce them are of different kinds, arising from quite different physical processes. The 21-centimeter radiation (more precisely 21.12 centimeters) of neutral hydrogen, for example, is generated by the quantum jump of the atom's single electron from one to the other of its two lowest-energy orbits. Most of the radiation from the Milky Way and from the discrete sources arrives, however, on a wide range of wavelengths. The typical short-wave map, at say 22 centimeters, charts the distribution of "thermal" radiation in the sky: radiation that is strong and roughly constant across the spectrum of short wavelengths and less intense at long wavelengths. The "nonthermal" radiation that is plotted in a typical long-wave map is bright at the long wavelengths and fades below the limits of detection at shorter wavelengths. As the term suggests, thermal radiation has its origin in the familiar process of excitation by heating. Stars, for example, may heat surrounding clouds of hydrogen to the temperature of ionization, separating the atoms into their constituent protons and electrons. These charged particles, accelerated by temperatures as high as 10,000 degrees absolute (degrees centi-

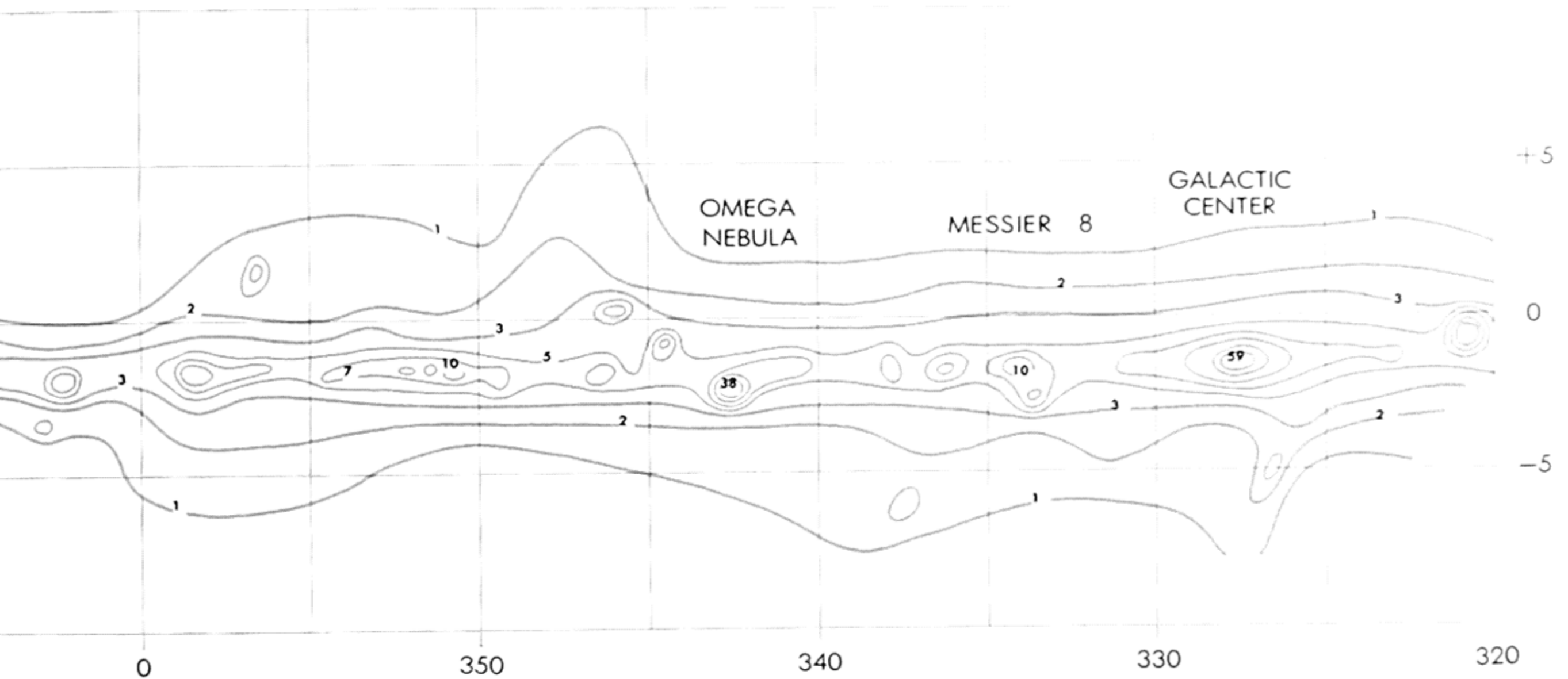
grade above absolute zero), move freely through the cloud. When an electron and a proton approach each other at high speed, they interact, causing each other to move in hyperbolic orbits and to lose energy [see *top illustration on page 428*]. They give up this energy as radiation, the wavelength depending upon their velocities and the distance at which they interact.

The origin of nonthermal radiation is not so well established, though we know that its intensity is too great to be explained by thermal processes. At present most of this radiation is attributed to the so-called synchrotron mechanism. In a synchrotron, electrons accelerated almost to the speed of light radiate on these wave-lengths (and others) when they are deflected into spiral paths around the lines of force in the machine's strong magnetic field. We know that our galaxy is permeated by magnetic fields, with their lines of force parallel to the galactic plane, and that the galaxy contains an as yet unknown quantity of ultrafast electrons. Presumably when the path of an electron is bent by one of these fields it, too, emits radiation.

Such theoretical understanding of the origin of the radiation at various wavelengths helped us derive the structure of the galaxy from radio maps. Another important set of clues is provided by matching the radio maps with photo-

graphic plates from such sources as the atlas compiled with the 48-inch Schmidt telescope on Palomar Mountain. We find that many of the radio sources at short wavelengths coincide with photographic images and, with almost equal significance, that almost none of the longwave sources do so. One whole set of coincidences shows us that most of the discrete sources in a short-wave map are emission nebulae, that is, diffuse clouds of hydrogen that surround bright stars and are highly excited by them. Both the light and radio emissions of such nebulae are given off by ionized hydrogen. The total radiation that they emit depends upon the size and the density of the gas clouds, both of which we can determine by radio-telescope measurements. Provided we have an optical determination of the distance, we can calculate the size of a nebula from its apparent diameter on a radio-telescope chart; and once its size is known, we can determine its density. Of course we can make the same estimates using light-brightness measurements, but they are more subject to error because the light from many nebulae is obscured by gas and dust clouds in intervening space.

Some emission nebulae are so completely hidden that only radio-telescopes can find them; the radio source near the galactic longitude of 358 degrees is a good example. This source is definitely a bright emission nebula, but no photo-



radio peaks. Most of these peaks, such as Messier 8, are emission nebulae: clouds of hot gas heated by stars embedded in them. Intense source on the right side of the map is the galactic center;

the bright source at the left is Cygnus A, which lies far outside our galaxy. Many radio sources in the Cygnus region are hidden by interstellar dust clouds which dim or completely cut off their light.

graphic counterpart of it has been found. We can thus assume that this nebula is completely concealed by interstellar clouds. But how can we be sure? Here knowledge of thermal radiation supplies the answer. Emission nebulae are thermal emitters, which means that their radiation is strong at short wavelengths. Once we know the intensity of their radiation at one wavelength, we know what radiation to expect at other wavelengths. Let us suppose, for example, that we pick up a new source at long wavelengths. We then tune the telescope to a wavelength in the short-wave end of the spectrum. If the object is an emission nebula, it will show up at this wavelength. If it is a nonthermal emitter, such as a cloud of ions trapped in an interstellar magnetic field, we will not pick up its radiation, since the radiation from most nonthermal emitters is too weak to detect at short wavelengths except with the help of the extremely sensitive receivers now being developed.

By comparing radiation maps of the galaxy made at the short wavelength of 22 centimeters with those made at longer wavelengths I have been able to estimate the amount of ionized hydrogen in the galaxy and plot its distribution. Since ionized hydrogen emits only thermal radiation, the first step was to subtract the nonthermal background. The 3.5-meter survey made by B. Y. Mills and his associates in Australia provided a

measurement of the nonthermal component of the total brightness of the Milky Way; it showed that almost half of the radio background at 22 centimeters in this region of the sky is nonthermal. From the separation of the thermal and nonthermal radiation it was possible to estimate the variation in thermal intensity along the ridge of the Milky Way and to measure the width of the thermal part of the ridge. The net result of this study was a reliable plot of the distribution of ionized hydrogen in the galaxy, most of it gathered into the numerous emission nebulae, some bright enough to be observed as discrete sources but many more too faint to be resolved.

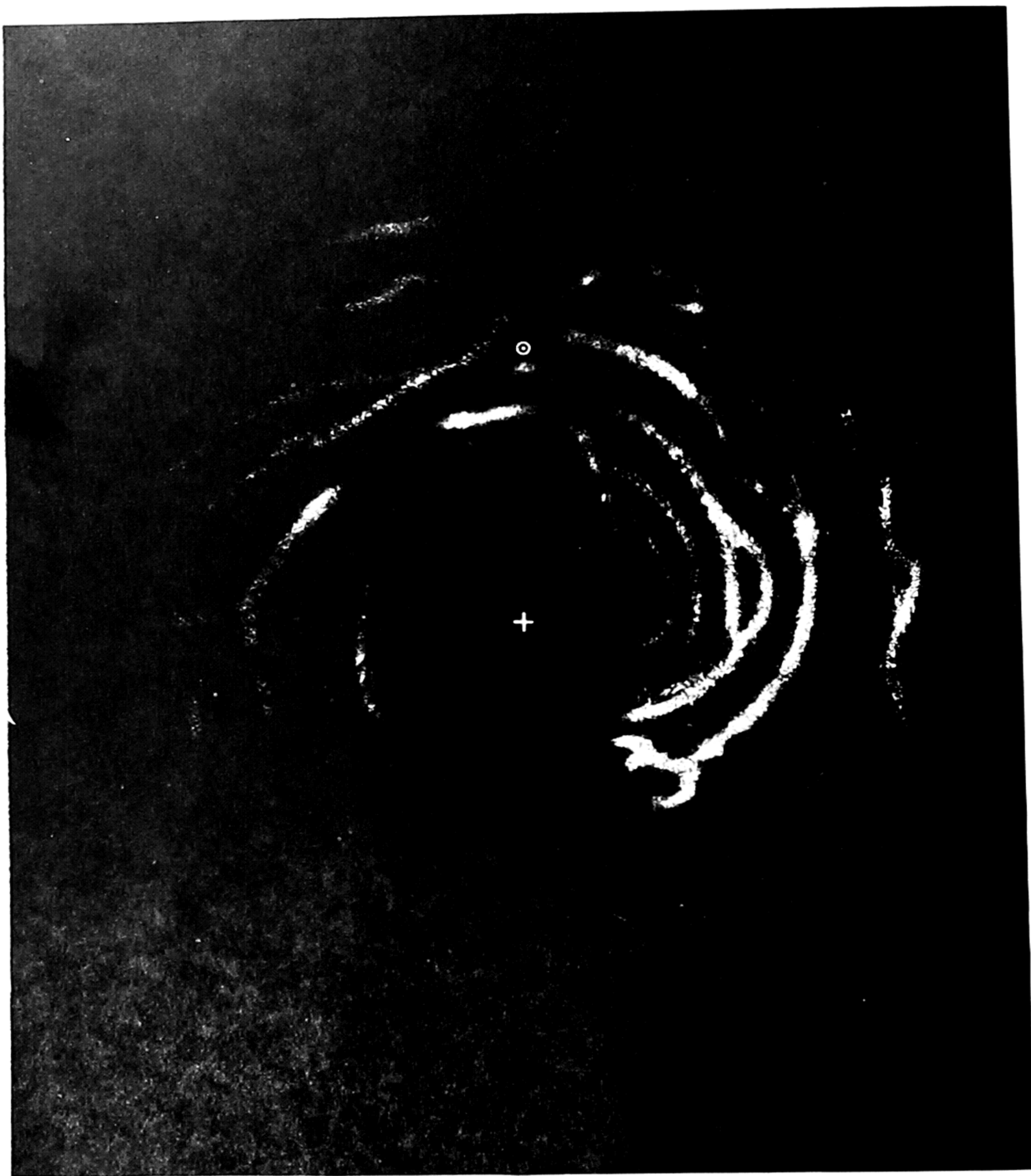
One fact emerged immediately. The ionized gas, along with the neutral gas and the stars, is concentrated in a very thin disk, not more than 600 light-years thick but over 60,000 light-years in diameter. The total mass of this disk is only a thousandth the total mass of the stars, dust and gas in the galaxy. By comparing radiation from ionized hydrogen at 22 centimeters to the 21-centimeter radiation from neutral hydrogen, we found that only 5 per cent of the hydrogen in the galaxy is ionized, a figure that agrees with earlier data gathered with optical spectrographs.

When we swing a radio telescope toward the center of the galaxy, we are able to see some crucial features of

its structure hitherto entirely concealed from observers. Between our sun and the center of the galaxy the density of ionized hydrogen slowly increases, reaching a maximum at about 14,000 light-years from the sun and 12,000 light-years from the center. Here, apparently, a ring of hot new stars ionizes the hydrogen around it. Beyond this ring, toward the center of the galaxy, the telescope shows that the density of ionized hydrogen drops off abruptly.

Well inside the ring of ionized hydrogen and new stars, at the very center of the galaxy, radio telescopes detect a hugely extended source of radiation, a region of nonthermal emission about two degrees wide in the plane of the galaxy and one degree deep at right angles to the plane. It encloses a few dense clouds of ionized hydrogen, with diameters of about .1 degree each. Here, apparently, is the nucleus of our galaxy. Similar blobs of light and radio brightness show up at the centers of nearby galaxies. No one can as yet explain the presence of these clouds at the center nor describe the mechanism by which they are ionized. From the rotational motion of the gas close to the center and from the intensity of the nonthermal emission it may be deduced theoretically that the nucleus contains stars and gas with a total mass equal to five million solar masses within a diameter of 100 to 200 light-years.

The Dwingeloo telescope tuned to 21



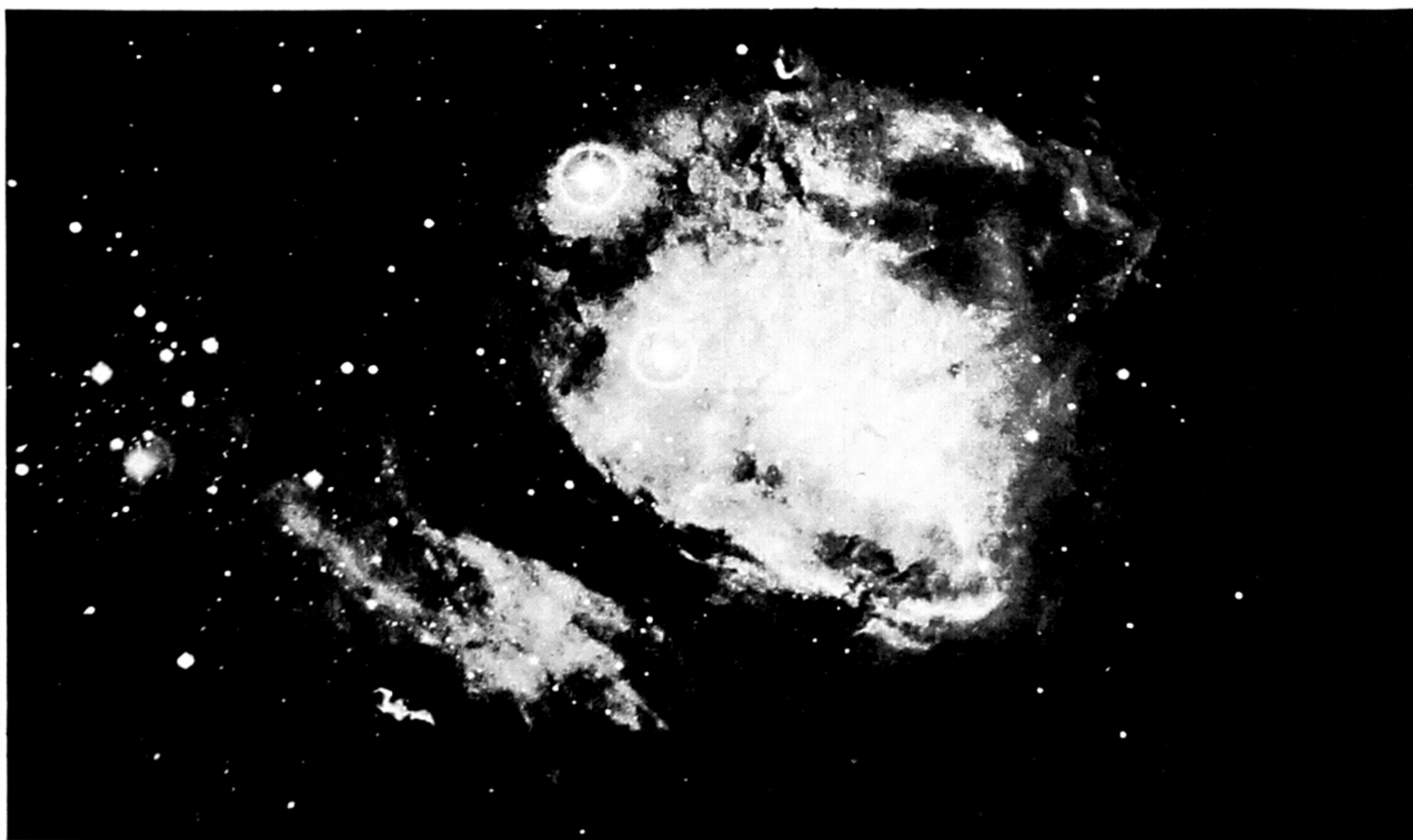
OUR GALAXY AS A SPIRAL NEBULA has been mapped during the past few years by Dutch and Australian radio astronomers, who have observed the 21-centimeter radiation from clouds of neutral hydrogen stretched along the spiral arms of the galaxy. Our earth and sun (*encircled dot*) are located about 27,000 light-years from the center of the galaxy (*cross*). The galaxy rotates around this center, our region completing its revolution in about 200 million years. We can detect this rotation by measuring the Doppler

shift in the radio signals from the drifts of hydrogen: If the gas moves toward us, its radiation shifts to slightly higher frequencies; if it moves away, its radiation appears at lower frequencies. We cannot detect the movement of the hydrogen clouds in the black, fan-shaped area extending from the sun to the opposite edge of the galaxy because it moves perpendicular to our line of sight, neither toward us nor away from us, and there is no detectable Doppler shift in its radiation. Hence this hydrogen does not appear on the map.



ONE OF THE BRIGHTEST RADIO SOURCES in the sky is Cygnus A, consisting of two colliding galaxies. Despite the intensity of their radio emission the galaxies are 270 million light-years away, and they appear only as two fuzzy spots in the middle of this photo-

graph made with the 200-inch telescope on Palomar Mountain. Although Cygnus A is the strongest source in the 22-centimeter map on the pages 424 and 425, both Cassiopeia A (the remnants of an exploded star) and the sun are brighter at this radio wavelength.



ANOTHER BRIGHT SOURCE, the emission nebula Messier 8, contains clouds of hot hydrogen that emit both radio waves and visible light. Some of the light is absorbed by the dark band of interstellar clouds that seems to divide the nebula. Messier 8 ap-

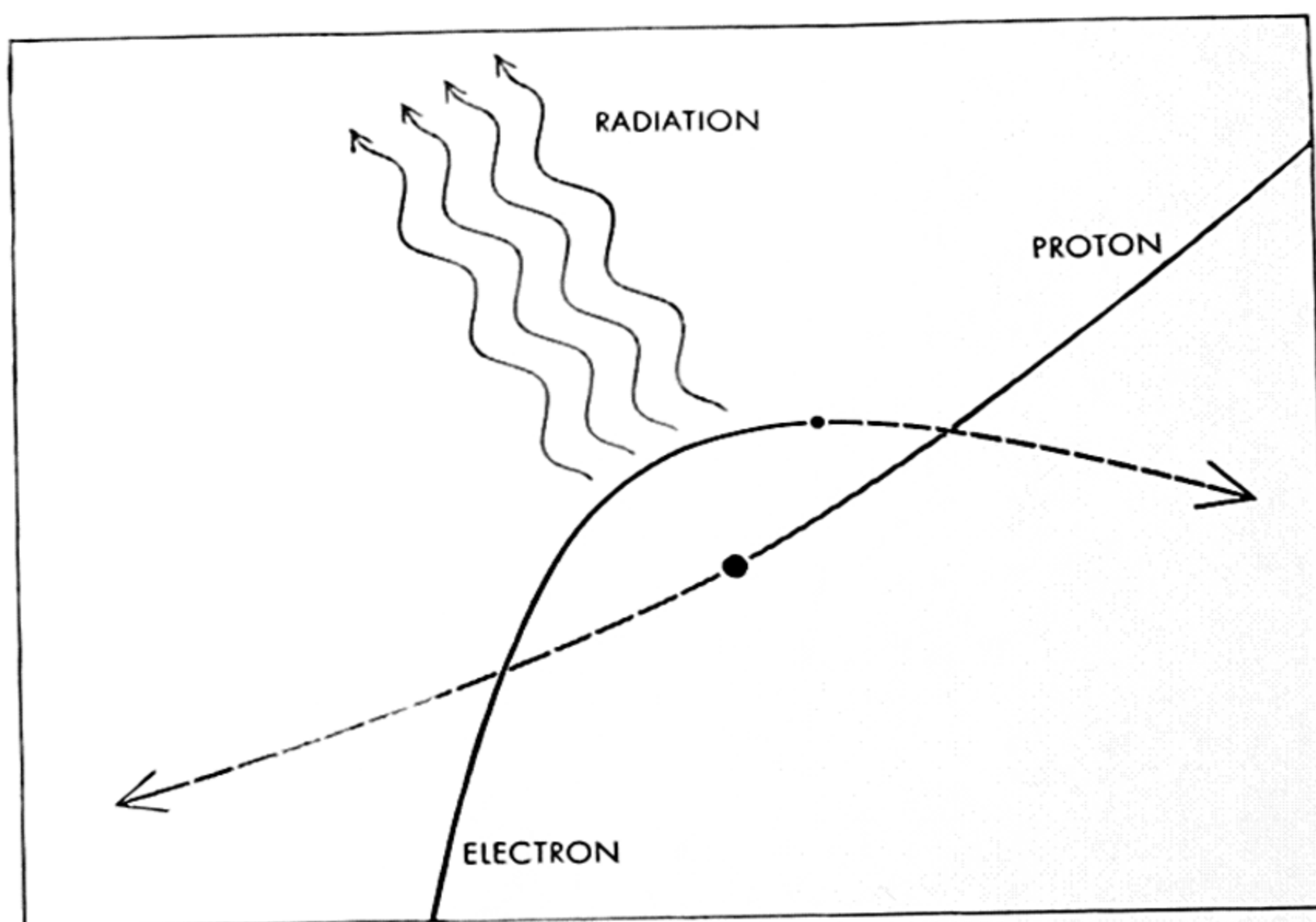
pears in the radio map on pages 424 and 425 as a small peak near a galactic longitude of 334 degrees; its radio signals pass unhampered through the dark clouds. Unlike the very distant source Cygnus A, the nebula Messier 8 lies within our own galaxy.

centimeters has revealed that neutral hydrogen surrounding the nucleus is moving radially outward from the center in the plane of the galactic disk at velocities varying from 50 to 200 kilometers per second. Apart from its radial motion the gas apparently also takes part in the general rotation of the galaxy under the influence of the galactic gravitational field. With the radial component of its velocity braked to 50 kilometers per second at about 10,000 light-years from the center, the density of the gas rises. It is just beyond this braking region, at 10,000 to 12,000 light-years from the galactic center, that the neutral hydrogen becomes ionized by the energy of hot new stars. Evidently the stars condense out of the gas; the higher density of the gas and the 10 million years it takes to flow through this zone provide favorable conditions for the formation of a large enough number of new stars to account for the ionization. Further study, with many more detailed observations, may lead to understanding of the evolutionary process by which a galaxy gives rise to new spiral arms.

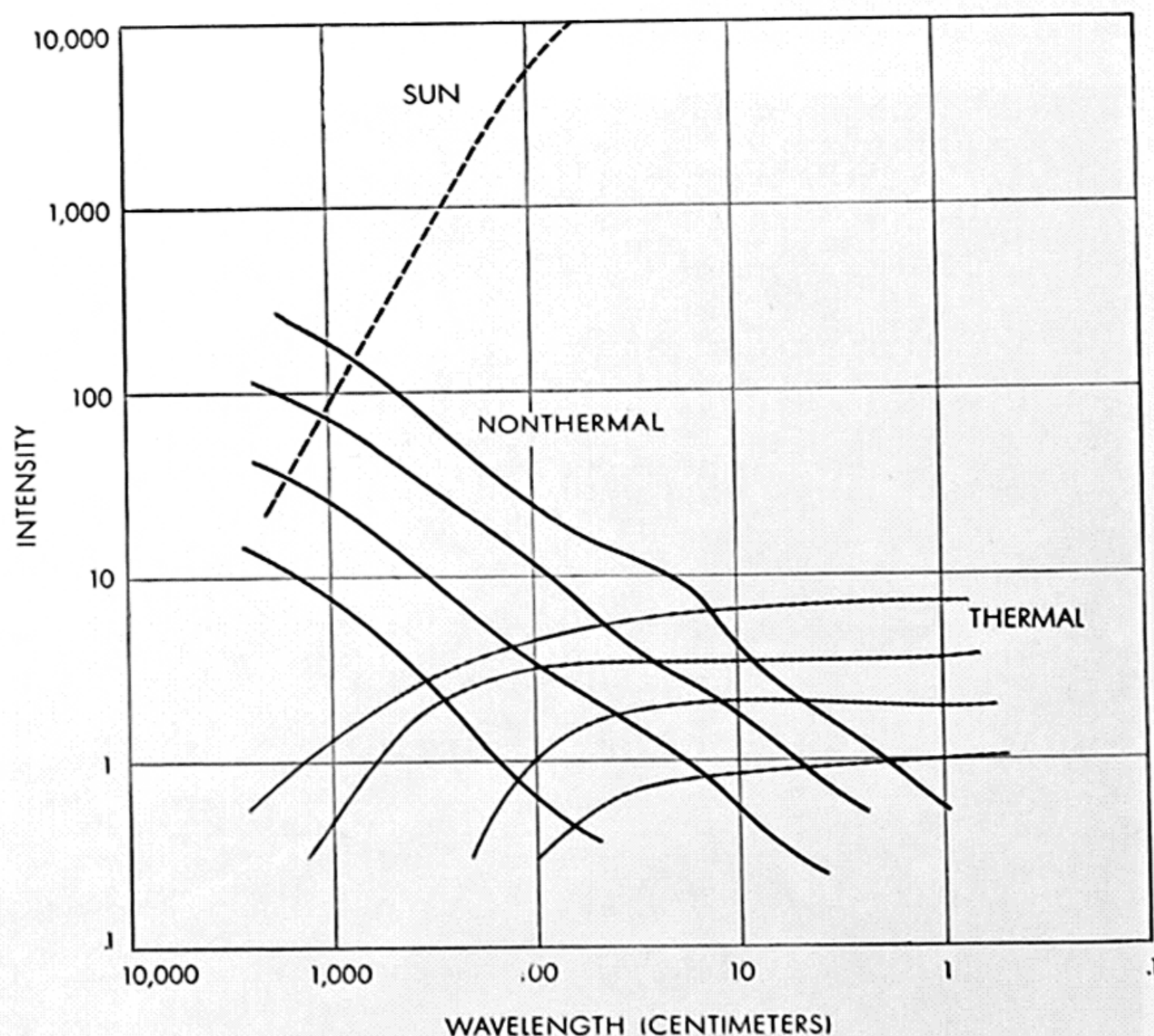
We do not, however, know where the neutral hydrogen that provides the substance of the new stars comes from. At the rate at which the hydrogen is emptying from the center of the galaxy, the entire supply must be exhausted in 30 million years, a day or two in the six-billion year history of the galaxy. Such a rapid loss of hydrogen would mean that the galaxy is exploding, a conclusion most astronomers are unwilling to accept.

It seems more likely that the supply of hydrogen at the center is continually replenished from some source outside. Perhaps this source is the galactic halo, a vast, thin sphere of ionized particles surrounding the galaxy which may be supplied in turn from hydrogen clouds ejected out of the galactic plane as the result of collisions or perhaps even from intergalactic space. We know almost nothing about the galactic halo, nor do we know by what mechanism, if any, its material is pulled unobserved into the center of the galaxy, although it is likely that magnetic forces would play some part in such a mechanism.

Even with very much better equipment it will be difficult to study the halo, because its density is extremely low. However, the telescopes and sensitive receivers needed to measure the distribution of ionized hydrogen in the galactic plane at wavelengths shorter than 22 centimeters will soon be available at various radio observatories in the U. S.



THERMAL RADIATION is produced when electrons and protons in ionized clouds of hydrogen pass each other at high velocities. The heavy proton is barely deflected from its path, but the lighter electron completely changes direction, thereby losing some of its energy. It gives up this energy by emitting light and radio waves. This radiation is called "thermal" because the original energy of the electrons depends on the temperature of the gas.



TWO MAIN CLASSES OF RADIO SOURCES, the thermal and the nonthermal emitters, can be distinguished from each other because thermal sources are weak at long wavelengths, but are strong and of approximately constant intensity at short wavelengths. Conversely, nonthermal sources are strong at long wavelengths and steadily diminish in intensity toward the short-wave end of the radio spectrum. The intensity of the sun is shown for comparison.

The Author

GART WESTERHOUT one of five or six full-time radio astronomers in the Netherlands, is in charge of radio astronomy at the Leiden Observatory. He was born 32 years ago in The Hague, and decided to become an astronomer at the age of 12, after attending lectures at the planetarium in The Hague. "By the time I was 14," he says, "I was lecturing friends and neighbors on astronomy." He studied at the University of Leiden under Jan H. Oort, now president of the International Astronomical Union, and H. C. van de Hulst, both pioneers in radio astronomy.

Westerhout has been doing research in that field since 1951, when the 21-centimeter emission line of neutral hydrogen was first detected. In 1958 he became the first astronomer in the Netherlands to receive a doctor's degree in radio astronomy.

Bibliography

THE CHANGING UNIVERSE. John Pfeiffer. Random House, 1956.

THE EXPLORATION OF SPACE BY RADIO. R. Hanbury Brown. John Wiley & Sons, Inc., 1958.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.



HUGE PROTON SYNCHROTRON under construction at Brookhaven National Laboratory is photographed from the air. Circular

tunnel housing its doughnut is 840 feet in diameter. This machine will produce particles of 25 to 30 billion electron volts (bev).

PARTICLE ACCELERATORS

by Robert R. Wilson

The evolution of these huge machines is still proceeding at a lively pace. They are used for two purposes: to "see" fundamental particles of matter and to create new ones.

From time to time in the course of history men have been swept up by intense currents of creative activity. In the pyramids of Egypt, in Greek sculpture and in Florentine painting we find monuments to such bursts of expression. My favorite example is the Gothic cathedrals that so magically sprang up in 12th- and 13th-century France, for I like to relate that magnificent preoccupation with construction to an obsession of our own time—the building of nuclear accelerators.

Like nuclear physics today, religion at that time was an intense intellectual activity. It seems to me that the designer of an accelerator is moved by much the same spirit which motivated the designer of a cathedral. The esthetic appeal of both structures is primarily technological. In the Gothic cathedral the appeal is primarily in the functionality of the ogival construction—the thrust and counter-thrust that is so vividly evident. So, too, in the accelerator we feel a technological esthetic—the spirality of the orbits of the particles, the balance of electrical and mechanical motion, the upward surge of forces and events until an ultimate of height is reached, this time in the energy of the particles. In both cases we find the architects working at the very limit of technical knowledge. In both there is intense competition between localities, regional and national. Both structures are expensive: a really large accelerator can cost \$100 million; the cost of a cathedral, in terms of medieval economics, was possibly higher.

But where a cathedral was a community enterprise, with many people in the region participating in its financing and construction, and nearly everyone in its enjoyment, an accelerator is esoteric. Its presence in a community is usually unknown and unsung. Few are the workers

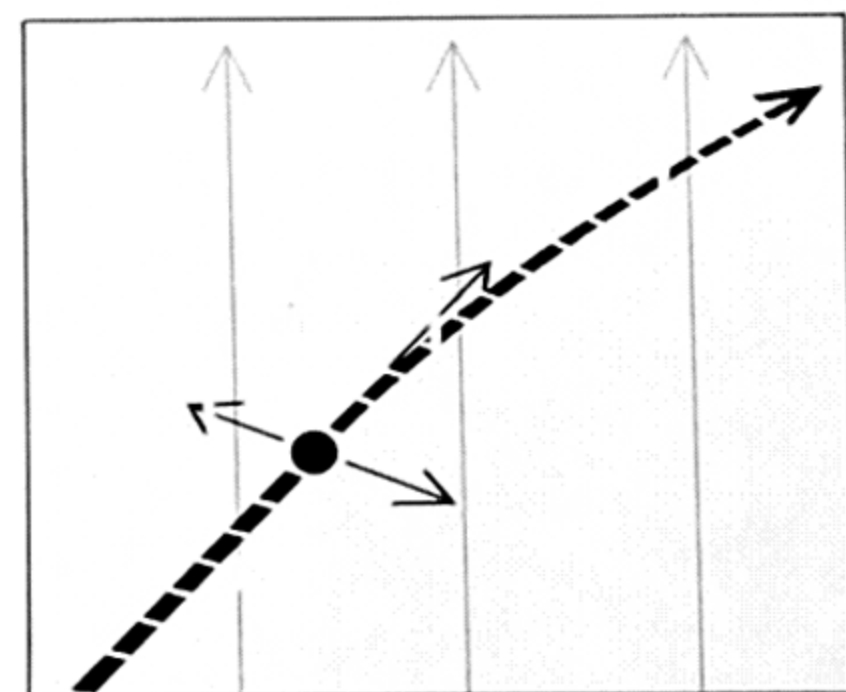
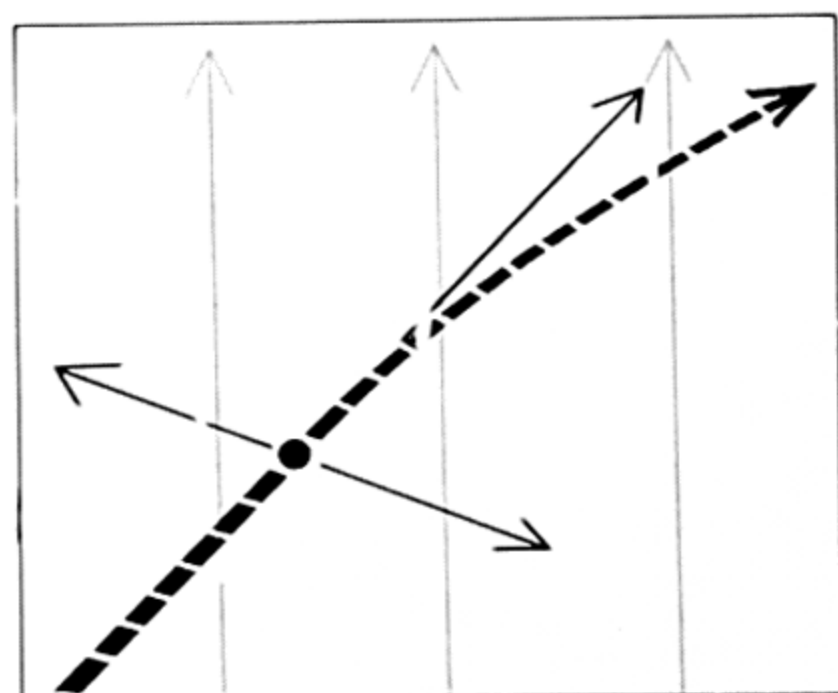
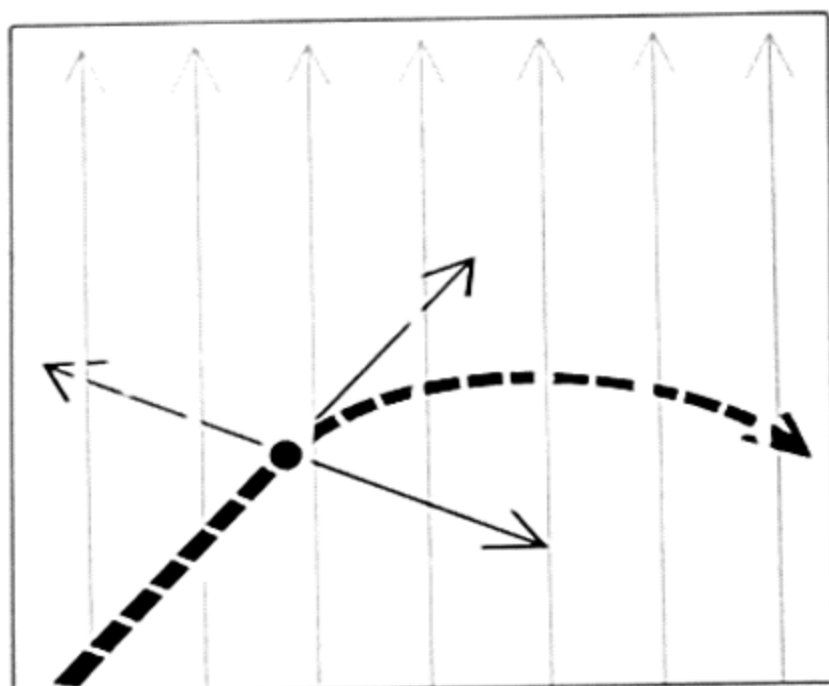
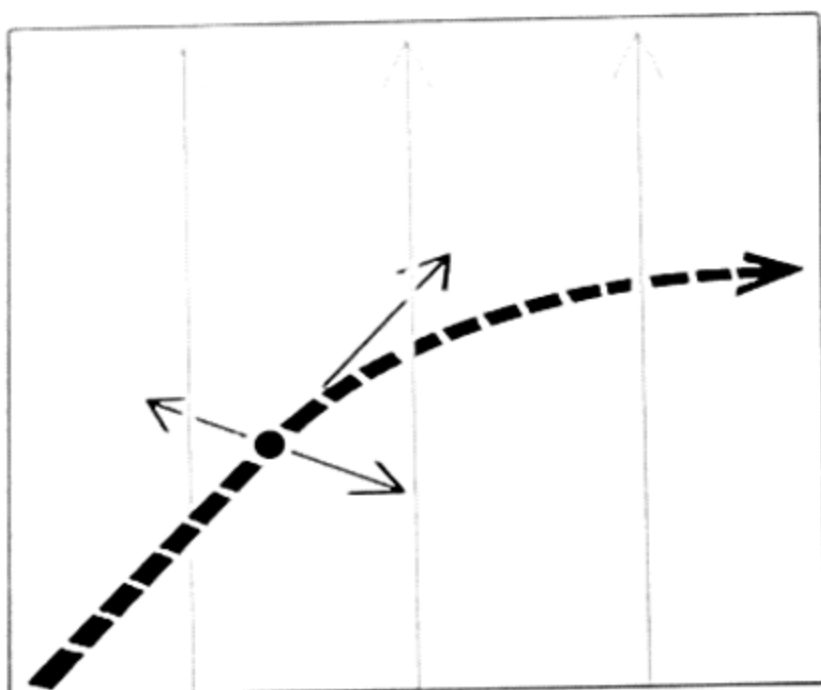
who help to build it, and fewer still are those who use it.

So the accelerator building boom goes on largely unnoticed, but at a quickening pace. Cyclotrons, the original "atom smashers," are now dotted almost all over the globe. They have evolved into

synchro-cyclotrons, and have reached their culmination in three giant machines, one at the University of California in Berkeley, another at the European Organization for Nuclear Research (CERN) in Switzerland and another in the U.S.S.R. These machines accelerate



PROTON SYNCHROTRON IN GENEVA is designed to yield 25 bev. Shown here is a section of the interior of its ring building. This structure is approximately 660 feet in diameter.



MAGNETIC FORCE on moving charged particles (black dots) is indicated by arrows pointing down and to right. Upward arrows show the speed of the particles and colored arrows the direction of the field. Large dot at the bottom represents a heavier particle.

protons to energies of between 600 and 700 million electron volts (mev). Synchrotrons, another development, are even bigger and more powerful. The Cosmotron, a 2,200-ton monster at Brookhaven National Laboratory which emits 3-billion-electron-volt (bev) protons, is small compared to the 6-bev, 10,000-ton Bevatron at Berkeley. This in turn is topped by the 10-bev, 36,000-ton Phasotron in the U.S.S.R. Two even larger machines are under construction at Brookhaven and CERN; they are designed to produce protons of 25 to 30 bev. And still bigger accelerators are being planned.

Nuclear Microscopes

Why? What is the purpose behind this almost feverish effort to build more and bigger machines? Perhaps the simplest answer is that accelerators are the microscopes of nuclear physics. We usually think of an accelerator as a sort of gun, producing high-speed particles which bombard the nucleus of the atom. But since particles are known to have wave properties, it is equally appropriate to say that the accelerator shines "light" on the nuclei, enabling us to "see" them.

Now the resolving power of a microscope, *i.e.*, its ability to distinguish small objects, depends on the wavelength of the light it employs. The shortest wavelength of visible light is about four 100,000ths (4×10^{-5}) of a centimeter; with these waves one can perceive a microbe of about the same length.

To examine smaller things, biologists now use the electron microscope. The wavelength of a particle depends on its mass and its energy. At a few thousand electron volts—the energy at which electron microscopes operate—an electron has a wavelength some 10,000 times shorter than that of visible light (about 10^{-9} centimeter). With these waves one can begin to see the details of molecules.

The nucleus of an atom is about 10^{-12} centimeter in diameter. This is the wavelength of a proton with an energy of 1 mev. To "see" the nucleus we therefore need a 1-mev proton "microscope," and to make out some of its internal details we need some 10 to 20 times as much energy. Thus a laboratory interested in classical nuclear physics will invariably have a Van de Graaff accelerator or a cyclotron operating in the range of 1 to 20 mev.

But physics has pushed beyond this point. At present many of us are interested not in the nucleus as a whole but in the structure of the protons and neu-

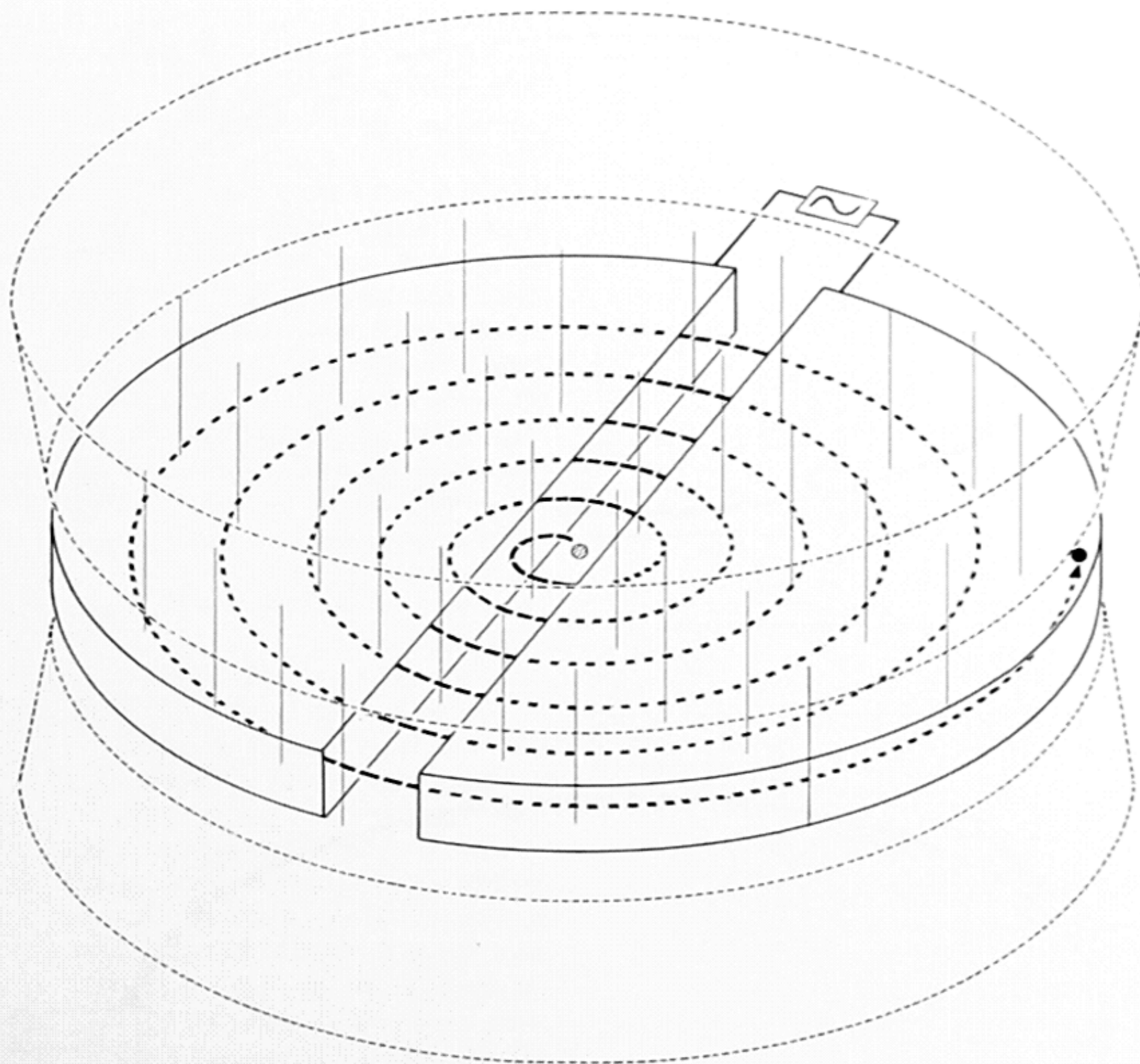
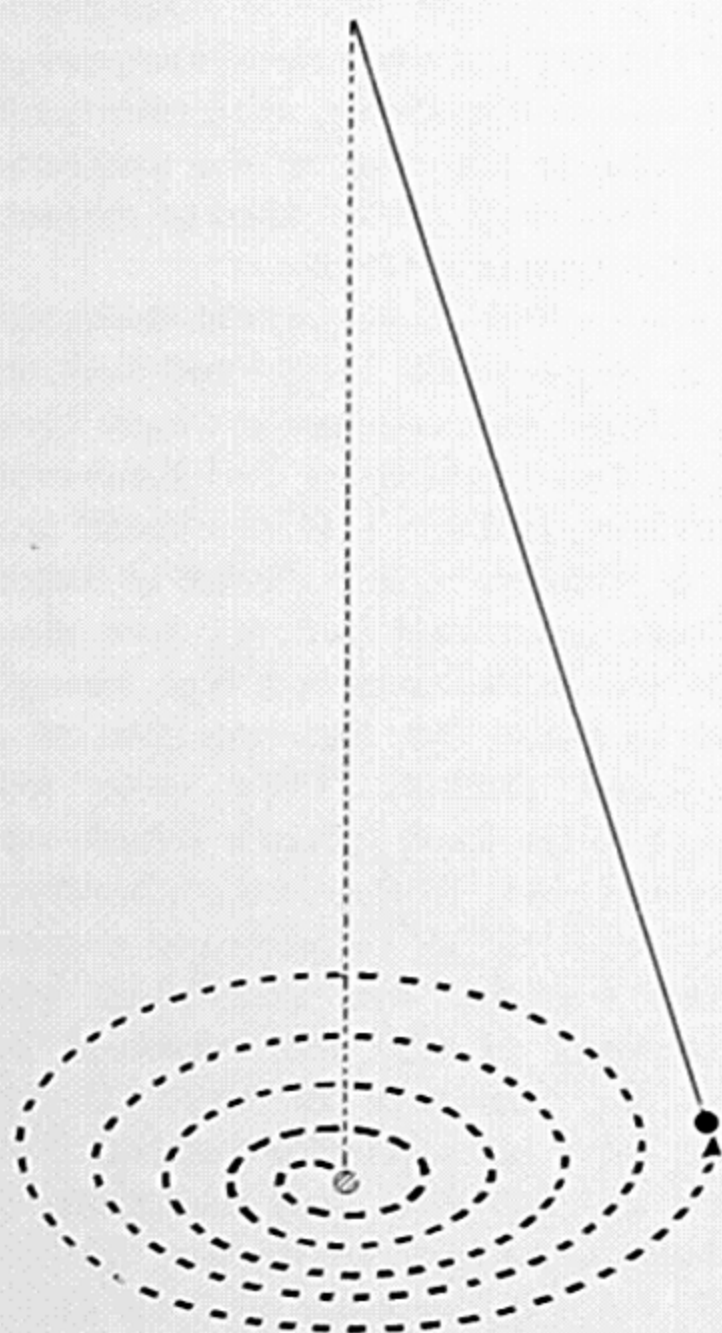
trons (nucleons) of which it is composed. It is the old problem of worlds within worlds, for the proton itself turns out to have a rich structure. It is perhaps 10^{-13} centimeter in diameter, and to resolve it requires an energy of several hundred mev. To see it in as fine detail as we can see the structure of the nucleus we must have still higher energy. It is for this reason that the 25- to 30-bev machines are under construction. If and when the structure of the proton is known, will its component parts turn out to have their own structure? Very possibly so, and if they do, machines of higher energy will be built to explore that structure.

The microscope analogy does not tell the whole story. When we get to sufficiently short wavelengths (*i.e.*, when the bombarding particles in our accelerators reach sufficiently high energy), we not only see particles, but we also make new ones. These new particles are created out of energy. At 1 mev an electron has enough energy to create a pair of particles—an electron and a positron. At 150 mev it makes pi mesons (pions) when it collides with a nucleon. Our 1-bev electron accelerator at Cornell University produces more massive particles: K and lambda mesons. The Bevatron, which produces 6-bev protons, is able to create antiprotons, antineutrons and still heavier particles such as xi and sigma mesons.

Thus as the energy of the machines has increased it has become possible to create more and heavier new particles. Obviously the exciting next step is to attain even higher energies, and then to see what sort of monster particles are created. One has the very strong feeling that new particles will indeed show up. It may well turn out that they will prove to be only complexes of particles which we already understand; however, it is exactly to answer such questions that we are building the machines.

Originally we constructed our accelerators in order to search for the ultimate in elementary particles. We expected these particles to be fragments and hence to be successively smaller; it was to improve our definition of them that we went to higher energies. Ironically the fragments now seem to get larger. One has the uneasy feeling that new machines make new particles which lead to the construction of new machines, and so on *ad infinitum*. In fact, there may be lurking here a new kind of indeterminacy principle which will inherently limit our knowledge of the very small.

So much for the reasons why accelera-



CYCLOTRON'S OPERATION is like that of a circular pendulum (*left*) in which the weight is pushed repeatedly to give an ever-widening swing. The schematic diagram at the right shows a particle (*dot*) spiraling within two D-shaped electrodes. The magnetic

pole pieces which provide the guiding field (*colored lines*) are outlined in light broken lines. The particles are accelerated by an oscillating electric field between the dees. The generator which produces the field is shown as a wavy line within a rectangle (*top*).

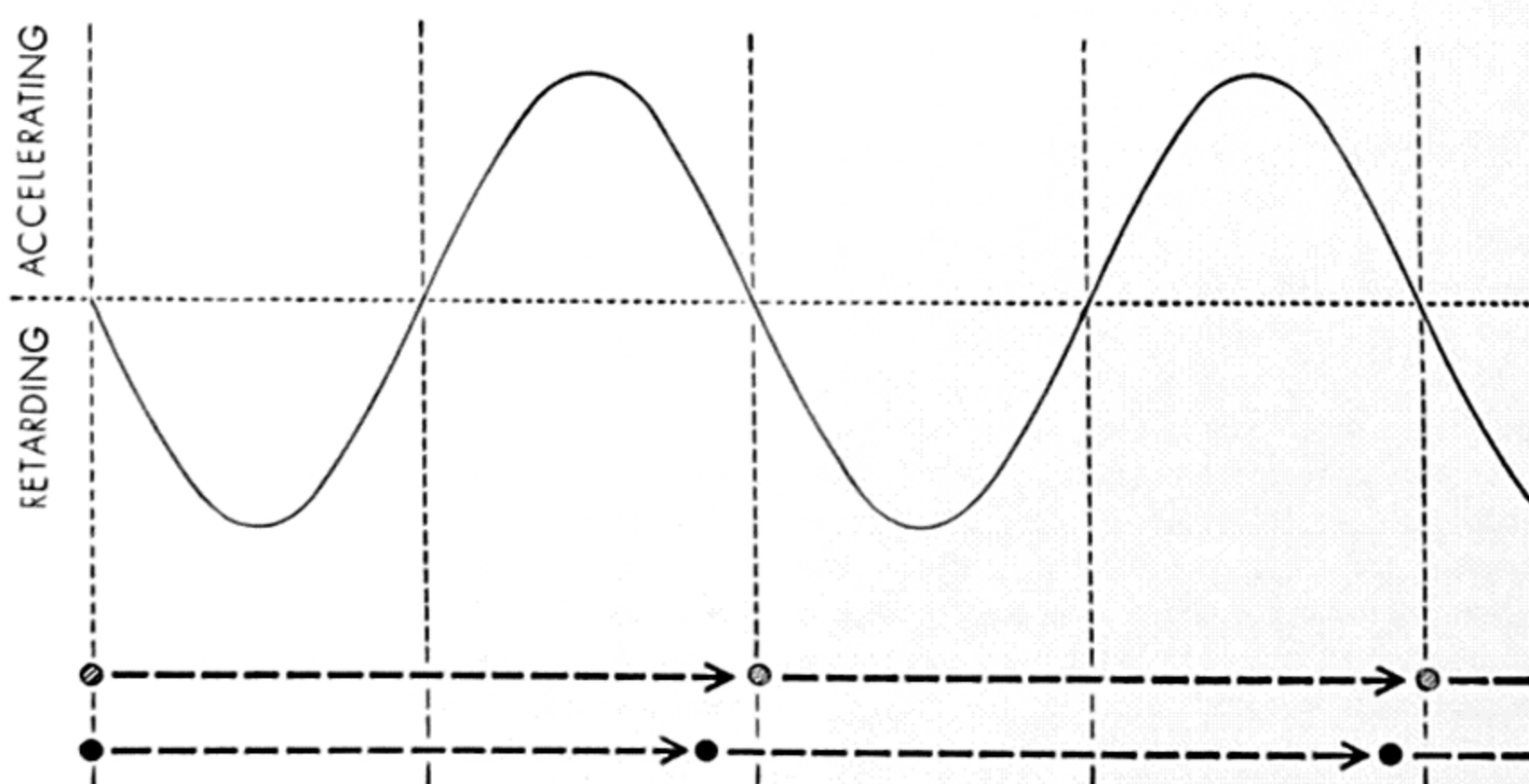
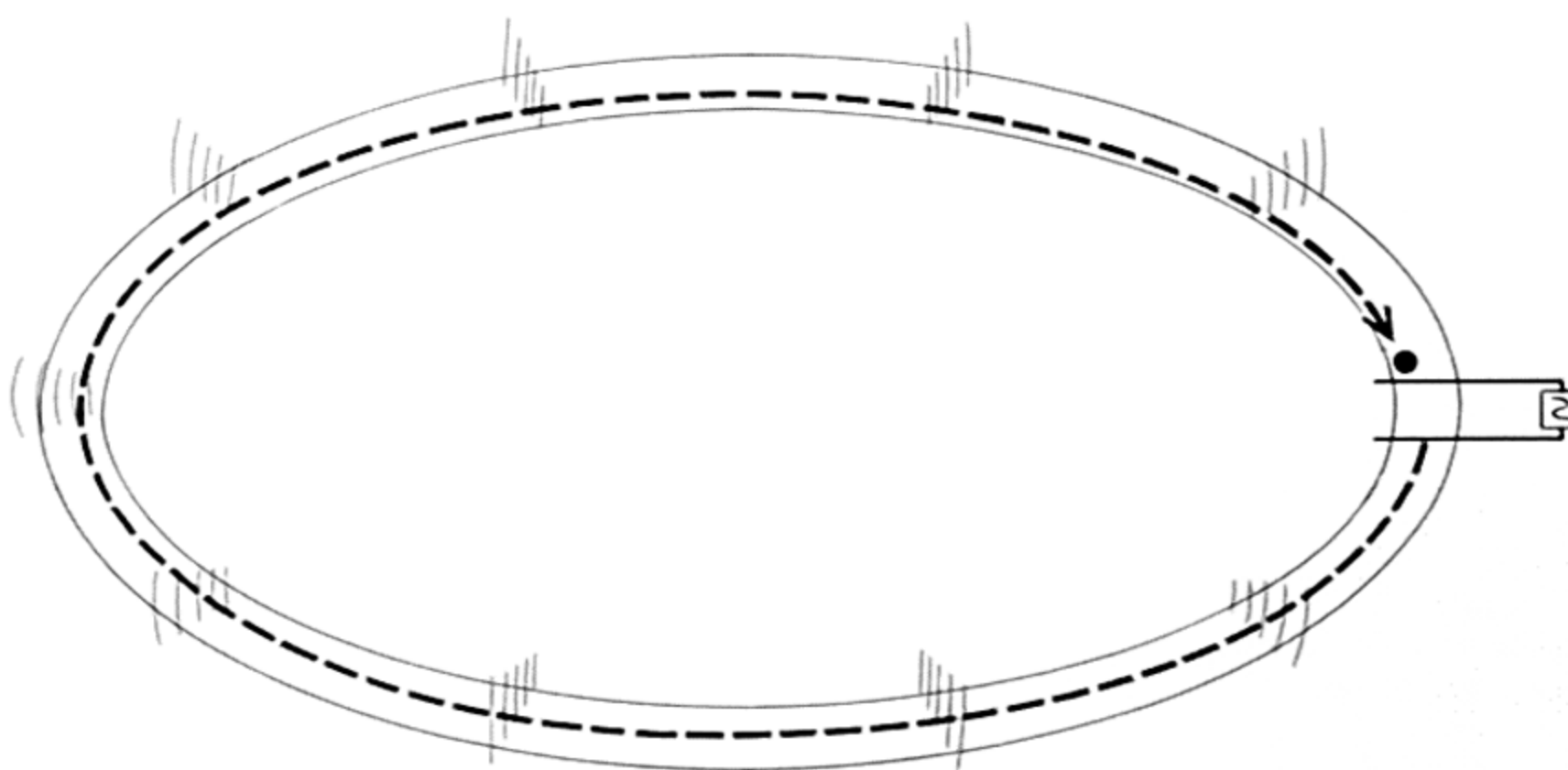
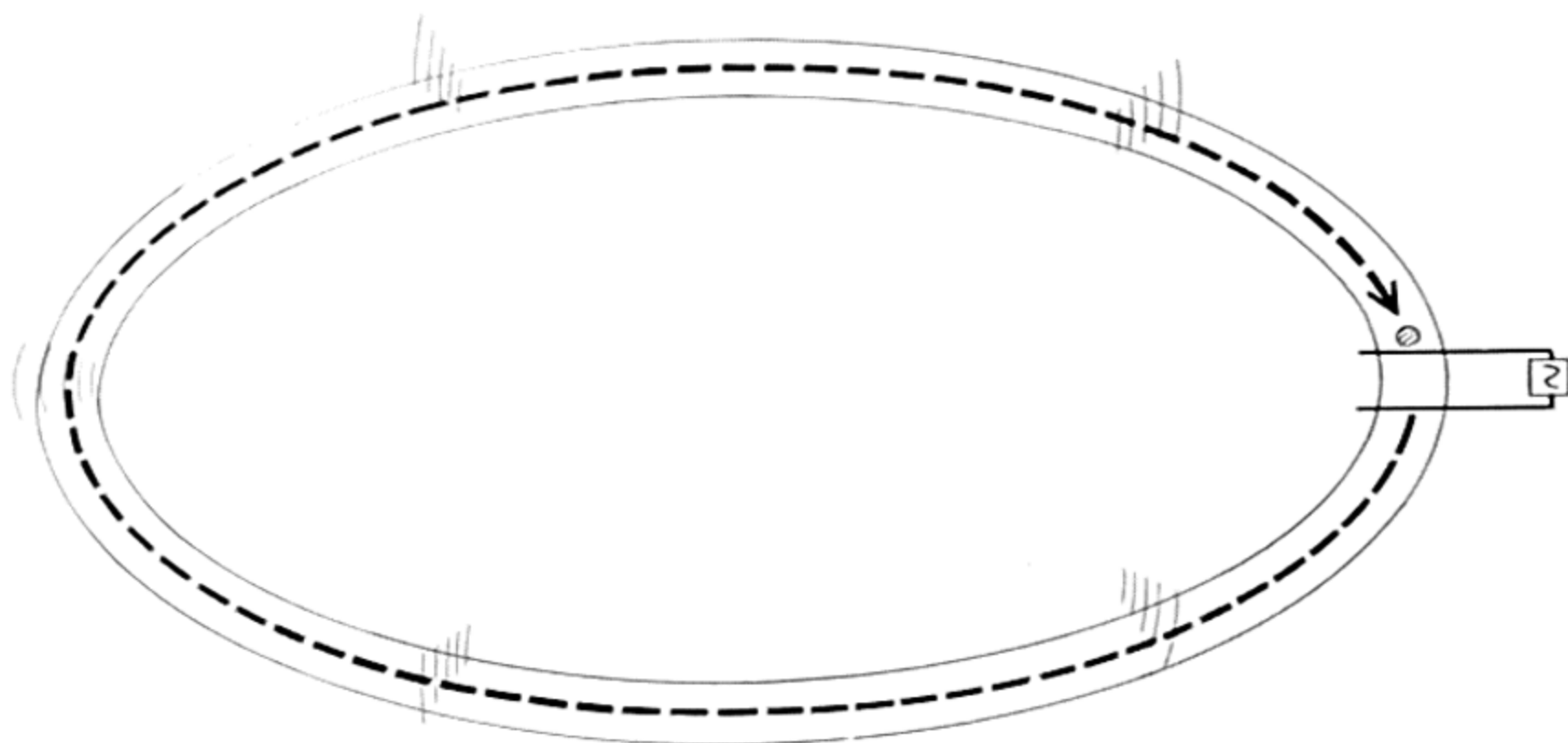
tors are built. Let us turn to the machines themselves. All of them operate on the same fundamental principle: charged particles (electrons or positive ions, usually protons) are put into an electric field which exerts a force on them, pushing them to high speeds and energies. (The electron volt, in which the energy is usually measured, is the energy acquired by a particle with one electronic unit of charge accelerated by a potential difference of one volt.) The simplest form of accelerator is a pipe along which a steady electric field accelerates the particles. This is the well-known Van de Graaff machine. To obtain higher energies a long pipe may be used with several accelerating electrodes which kick the particles to higher and higher speeds as they travel down the tube [see "The Linear Accelerator," by Wolfgang Panofsky; *SCIENTIFIC AMERICAN* Offprint 234]. But to attain a really high energy by this method would require an extremely long pipe. To get around this difficulty the particles can be made to travel in a circular or spiral

path which brings them back through the same electrodes where the accelerating voltage is applied again and again.

It is with such circular machines that we are chiefly concerned in this article. In these machines the circular motion is brought about by magnetic fields. A magnetic field exerts a force on all electric charges that move through it; the force is always at right angles to the direction of the charges' travel. It is the same kind of force that acts on a stone whirled at the end of a string. The magnetic field, like the string, forces the particles to move in a circular path. The stronger the field, the sharper the curvature of the path; on the other hand, the faster or heavier the particle, the less it is curved by a given field [see *diagrams on opposite page*].

The simplest and oldest type of accelerator to make use of magnetic bending is the cyclotron. The operation of this machine can be most easily visualized by imagining a weight suspended by a string and pushed so as to describe a circular motion. As with any pendulum the

time required to complete a full circular swing is the same whether the circle is small or large. Thus if the weight is pushed rhythmically it will move outward in an ever-widening circle, returning to the pushing point in the same time on each revolution [see *diagram above*]. So it is in the cyclotron: each ion whirls inside of two semicircular electrodes or "dees," getting an electrical push when it passes from one to the other. A vertical magnetic field provides a constant inward push and, like the string, holds the ion in a circular path and guides it back to the gap between the dees, where it is given another electrical push. The velocity of the ion then becomes greater and, as a result of its inertia, the curvature of the circular path caused by the magnetic field becomes larger. The time taken to traverse a full circle is the same no matter how big the radius, because the increase in speed just compensates for the increase in path-length per turn. Now if the voltage across the dees is made to oscillate rapidly, and if its period is adjusted so that it exactly matches



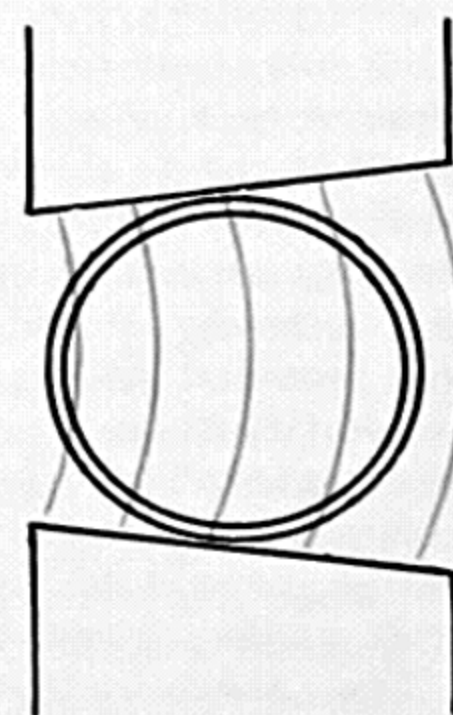
SYNCHROTRON restricts particles to a nearly circular path by means of a magnetic field (*colored lines*) which grows stronger as the particle energy increases. At top an electron (*hatched circle*) is in an orbit that brings it to the accelerating gap (*right*) just as the voltage changes from accelerating to retarding (*curve at bot-*

tom). In the center drawing the field is made stronger and the electron (*black circle*) is bent more strongly, following a shorter path and arriving at the gap in time to get a push. After a number of pushes it spirals out to the original path. The cross section at bottom right shows magnetic pole pieces around the doughnut.

the period of revolution of the ions, then the ions will be pushed in the right direction at the right time at each crossing of the gap between dees; the energy of the ions will build up until their path takes them to the edge of the magnetic field, where they can be used or extracted in the form of a beam.

If the cathedrals had great designers such as Suger of St. Denis and Sully of Notre Dame, the accelerators have their Cockcroft of Cambridge and Lawrence of Berkeley. In 1928 J. D. Cockcroft and E. T. S. Walton built a device in which a voltage generated between two electrodes accelerated ions to a high enough speed to cause the disintegration of a bombarded nucleus. They were still working in the magnificently simple tradition of Ernest Rutherford's laboratory at the University of Cambridge. A quite different tradition was established with the building of the first cyclotron by Ernest O. Lawrence in 1930. It has spread from his laboratory at the University of California and has come to dominate experimental nuclear physics in this country. Indeed, one can begin now to trace this spirit abroad, particularly to the U.S.S.R., where it may flourish even more vigorously than it does in the U. S.

This tradition, called "berkelitis" by its detractors, is a true departure in experimental physics. Previously experimental equipment had been constructed to test a particular surmise or idea. But building a large accelerator is more analogous to outfitting a ship for an expedition of exploration, or to the construction of a huge telescope to study a variety of astronomical objects. After several cyclotrons had been built at Berkeley, the



students and associates of Lawrence traveled far and wide to spread the gospel. By World War II they had helped to build cyclotrons not only at universities in the U. S., but also in several other countries. The biggest of these machines produced protons of about 10 mev. As we have seen, this is an appropriate energy for exploring the nucleus as a whole, but not for examining its parts. Just before the war Lawrence had begun to build a giant cyclotron, to enter the energy region above 100 mev, with which he could start to probe nucleons.

The Synchrotron

It was characteristic of Lawrence that he went ahead despite a prevalent conviction that the energy limit of the cyclotron was about 20 mev. This conviction was based on an effect predicted by Albert Einstein's theory of relativity: particles traveling at nearly the speed of light will increase in mass. At 20 mev a proton has entered this "relativistic" region, and further increases in energy will result not so much in greater speed as in greater mass. When this happens, the particle in a cyclotron begins to fall behind schedule as it spirals farther outward, and it no longer arrives between the dees at the right time to get a push from the oscillating voltage.

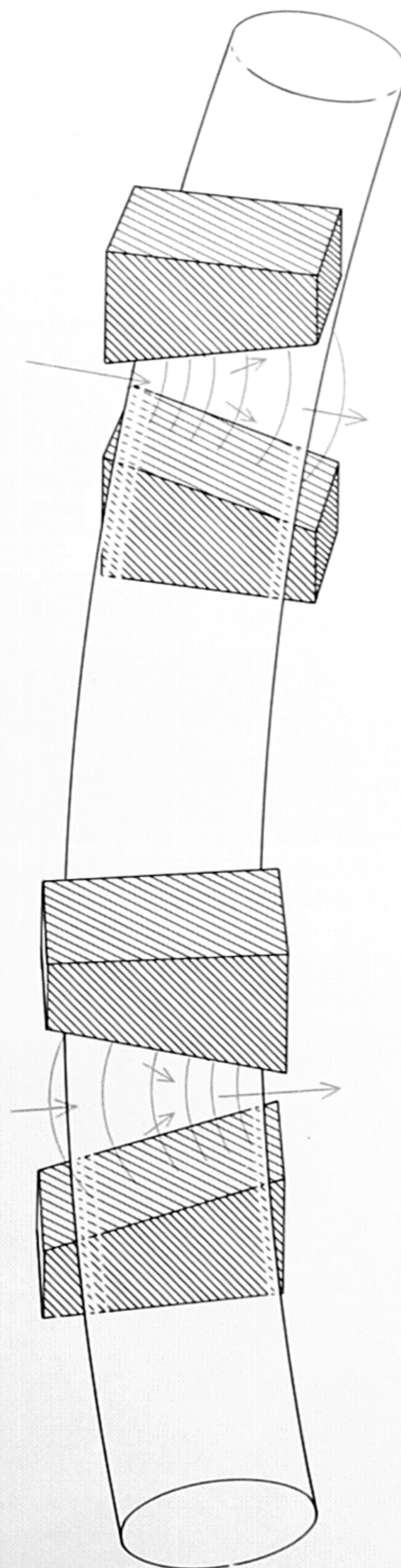
The war interrupted work on Lawrence's big machine. Its huge magnet was used to separate isotopes of uranium for the atomic-bomb program. At the end of the war V. I. Veksler of the U.S.S.R. and E. M. McMillan of the University of California independently and almost simultaneously enunciated the so-called synchrotron principle. This principle showed the way to accelerating particles into the completely relativistic region. It was exactly the sort of *deus ex machina* that Lawrence had envisioned when he gambled some \$1 million in starting his big cyclotron. The principle was immediately adopted. A successful synchro-cyclotron was built which produced protons in the region of 100 mev (eventually 730 mev). In the next few months a number of important features of the proton were discovered.

To understand the synchrotron principle, it is easier to consider its application in the electron synchrotron rather than in the more complicated synchro-cyclotron. Some half-dozen of these electron accelerators, with maximum energies of about 300 mev, were also built just after the war.

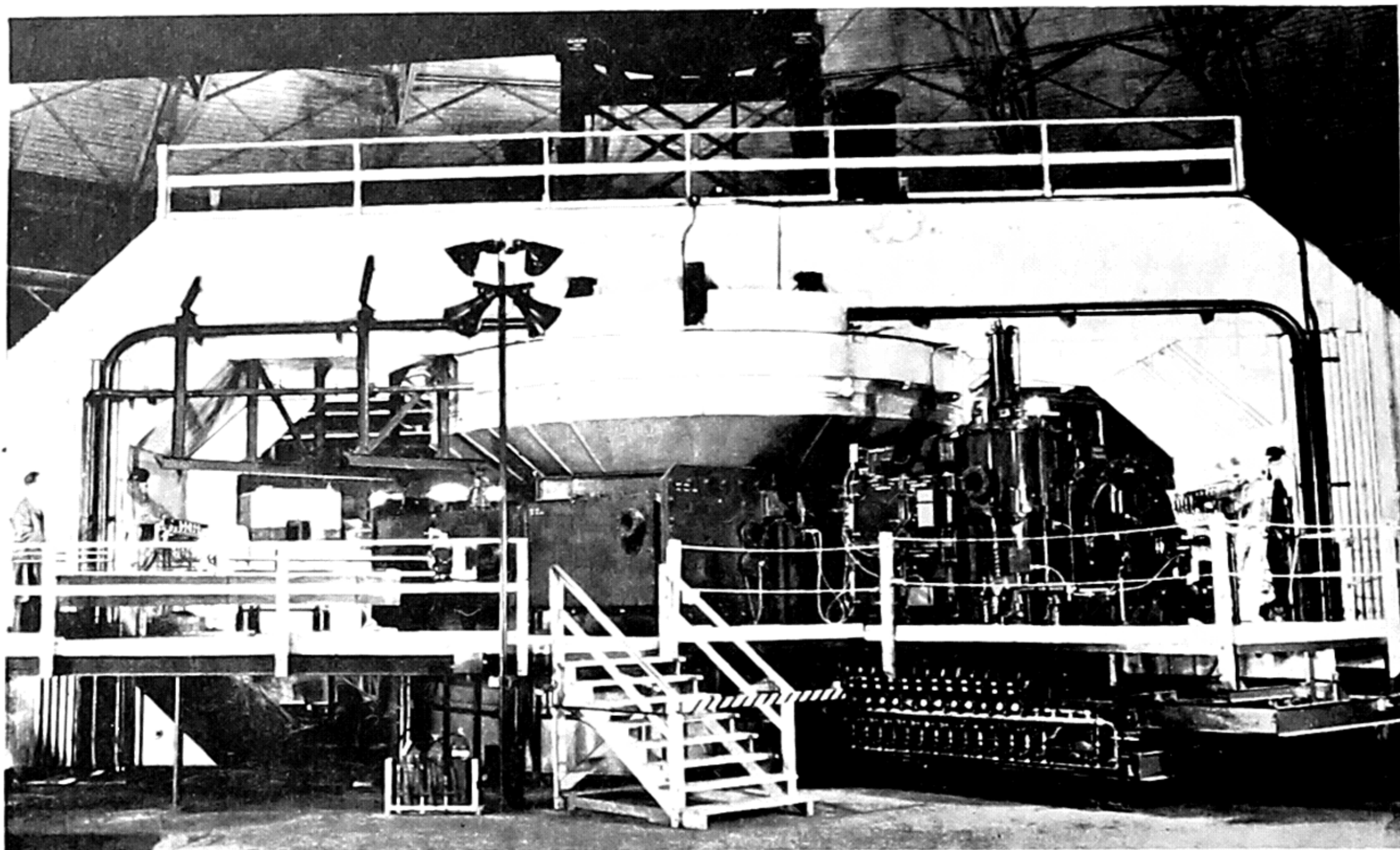
In a synchrotron electrons travel on a circular orbit inside a narrow doughnut-

shaped vacuum vessel. At one point in the doughnut is a pair of accelerating electrodes across which there is an oscillating voltage like that in the cyclotron. A ring-shaped magnet surrounding the doughnut produces a field which forces the particle to travel on orbits close to the center of the tube [see diagram on opposite page]. The electrons are injected into the doughnut from a small linear accelerator at an energy of about 2 mev. At this energy their speed is some 98 per cent of the speed of light; hence they cannot travel much faster. To make matters simpler let us assume that the speed is exactly the speed of light and that the whole increase in energy goes into mass. Now imagine an electron in a circular orbit at the center of the doughnut. The electron is held there by a constant magnetic field. Also imagine that our oscillating voltage is applied, but that the electron crosses the accelerating gap just at the time when the voltage falls through its zero value. The frequency of the voltage is made the same as that of the electron traveling around its orbit at the constant speed of light. The electron now passes the gap on all subsequent turns just as the voltage becomes zero. Thus nothing happens; the electron remains on its orbit and keeps the same energy. Now we increase the magnetic field slightly. Since the energy (mass) is still the same, the particle is forced into a sharper curve, *i.e.*, its orbit gets smaller. But because the orbit is smaller and the speed is constant, the time it takes the electron to return to the accelerating gap is shorter. Hence the electron arrives slightly before the voltage has fallen to zero; it is accelerated slightly. On the next turn, if the energy is still not large enough, the orbit will still be too small: the electron will arrive still earlier and be accelerated even more. Eventually the energy will increase enough (that is, the electron will get heavy enough) so that it is bent less sharply and edges out to its original orbit. If the energy should become too great, the orbit will be too big and the time it takes the electron to make each turn will be too long. This will cause the electron to drop behind the accelerating voltage and be pushed backward so that it will lose energy. Thus we have a beautiful automatic device for keeping the electron on the right orbit, or at least oscillating around the right orbit. That is all there is to the synchrotron principle or, as it is sometimes called, phase focusing.

Now we can see that, if the magnetic field of the synchrotron is increased continuously, the energy of the electrons

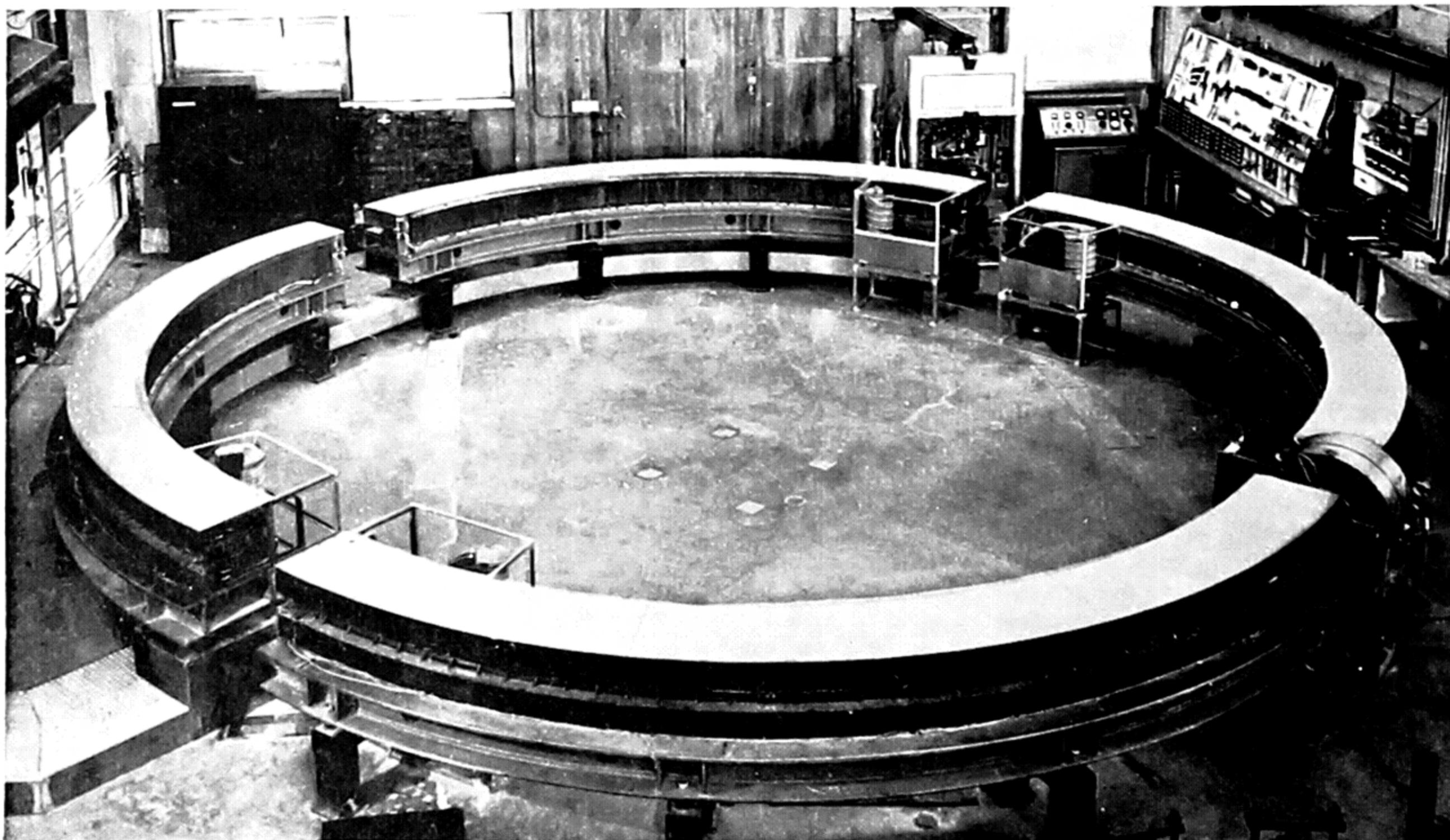


STRONG FOCUSING is produced by magnetic fields which are alternately bowed out and in. Horizontal arrows show radial forces on the particles at inner and outer edges of the field. Slanted arrows represent forces which focus or defocus particles vertically.



SYNCHRO-CYCLOTRON at the Berkeley Radiation Laboratory of the University of California is now the most powerful machine

of its kind. A modification of its design last year increased the energy of its proton beam to 730 million electron volts (mev).



ELECTRON SYNCHROTRON was photographed in the author's laboratory at Cornell University while its guiding magnet was un-

der construction. Machine, which produces an energy of 1 bev, is the first to use strong focusing. Accelerating electrodes are at right.

will also increase continuously; the electrons will receive energy at just the right rate to keep them on the central, or synchronous, orbit. In practice electrons can be injected into the doughnut when the magnetic field is rather weak (about 10 gauss) and ejected when the field is quite strong (more than 10,000 gauss). A synchrotron with a large enough radius can accelerate electrons up to energies of about 10 bev. There are now about six machines, built or being built, which are designed to yield electron energies between 1 and 1.5 bev. At Cambridge, Mass., a 6-bev electron synchrotron is being constructed by a joint Harvard University-Massachusetts Institute of Technology group.

Let us return to the synchro-cyclotron. It works in essentially the same way as a synchrotron but it is shaped like a cyclotron. Instead of a varying magnetic field it has a constant field, but the frequency of the accelerating voltage applied to the dees is varied. This means that the synchronous orbit of the protons is not a fixed circle but a growing spiral.

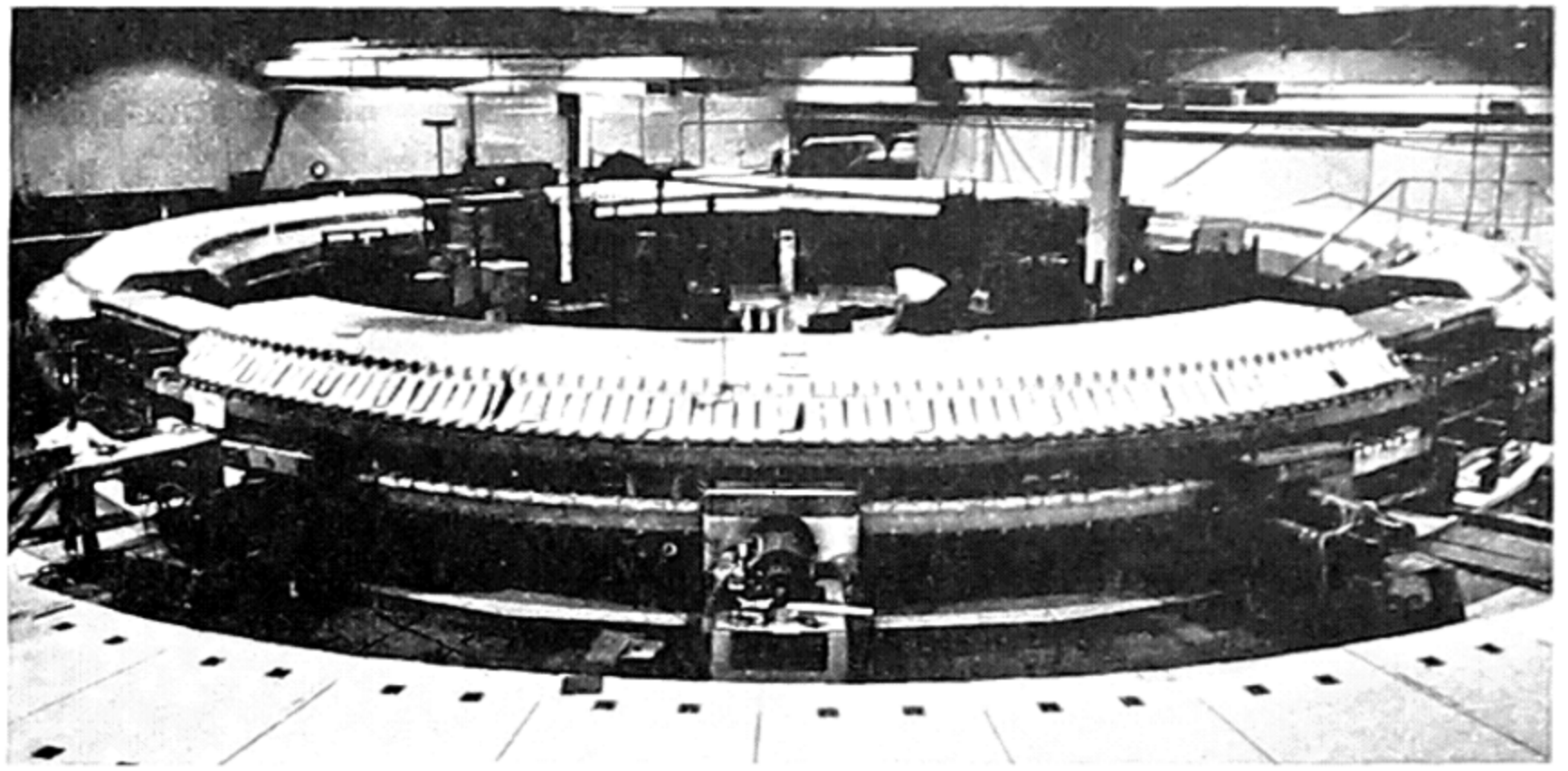
In another class of accelerators, the proton synchrotrons, both the magnetic field and the frequency of the accelerating voltage are varied. The increasing field counteracts the protons' tendency to spiral outward as they get up to relativistic energies, and the orbit is again a fixed circle. Above about 5 bev the protons are traveling practically at the speed of light, and from here on the proton synchrotron works just like an electron synchrotron.

If I may extend the figure of speech with which I began this article, each kind of accelerator has its own architectural style. To me synchro-cyclotrons are baroque. Proton synchrotrons are definitely Romanesque, although their rounded arches are horizontal. Electron synchrotrons have a lightness and grace that could only be Gothic.

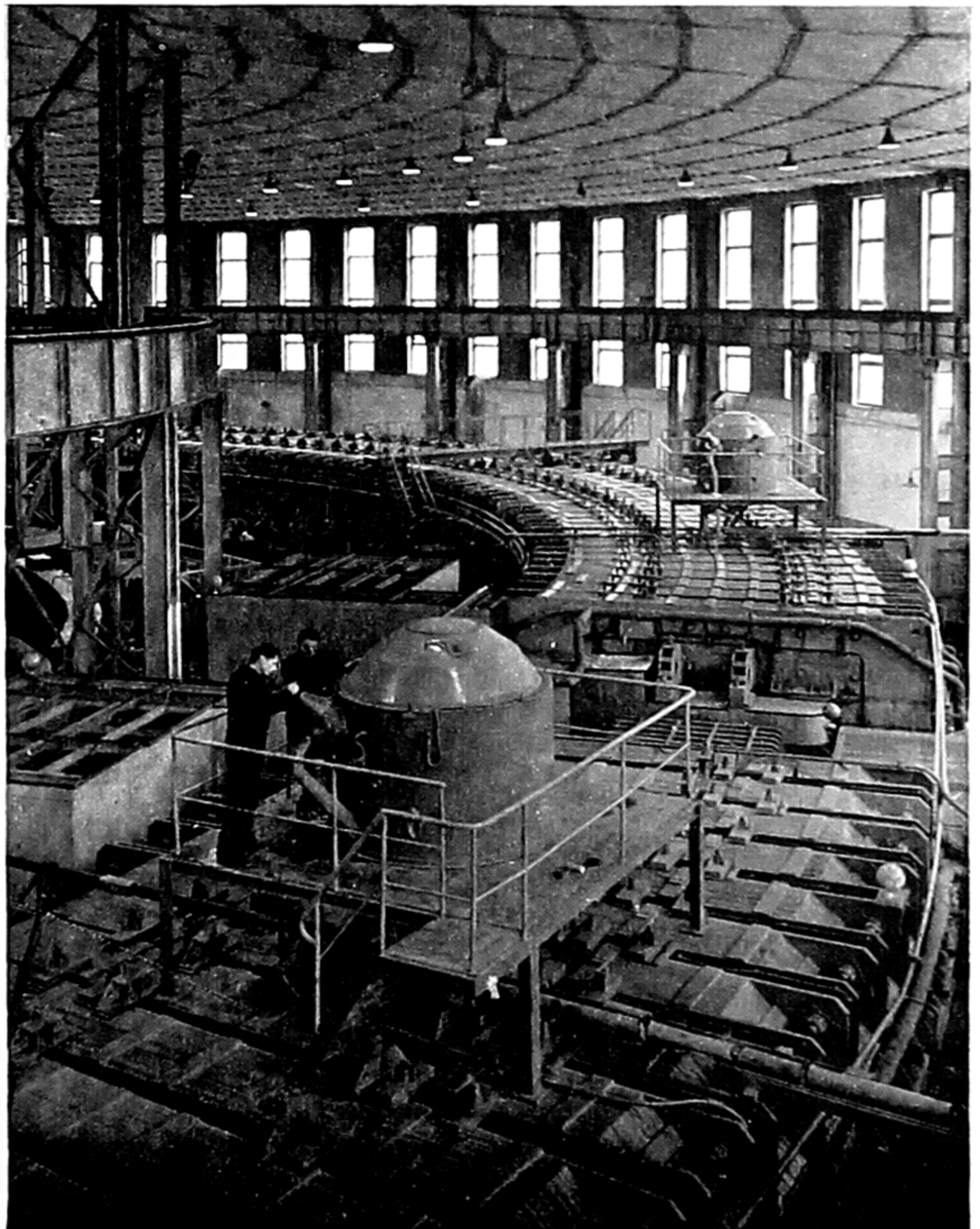
The Newer Machines

This brings us more or less up to date in the evolution of accelerators. We may now ask whether we are near the end of this movement in physics or still at its beginning. The field still has tremendous vigor, and it is my guess that we are at about the same stage as the cathedral builders were after they had completed Notre Dame of Paris. The significant innovations were behind them, but most of their masterpieces were yet to come.

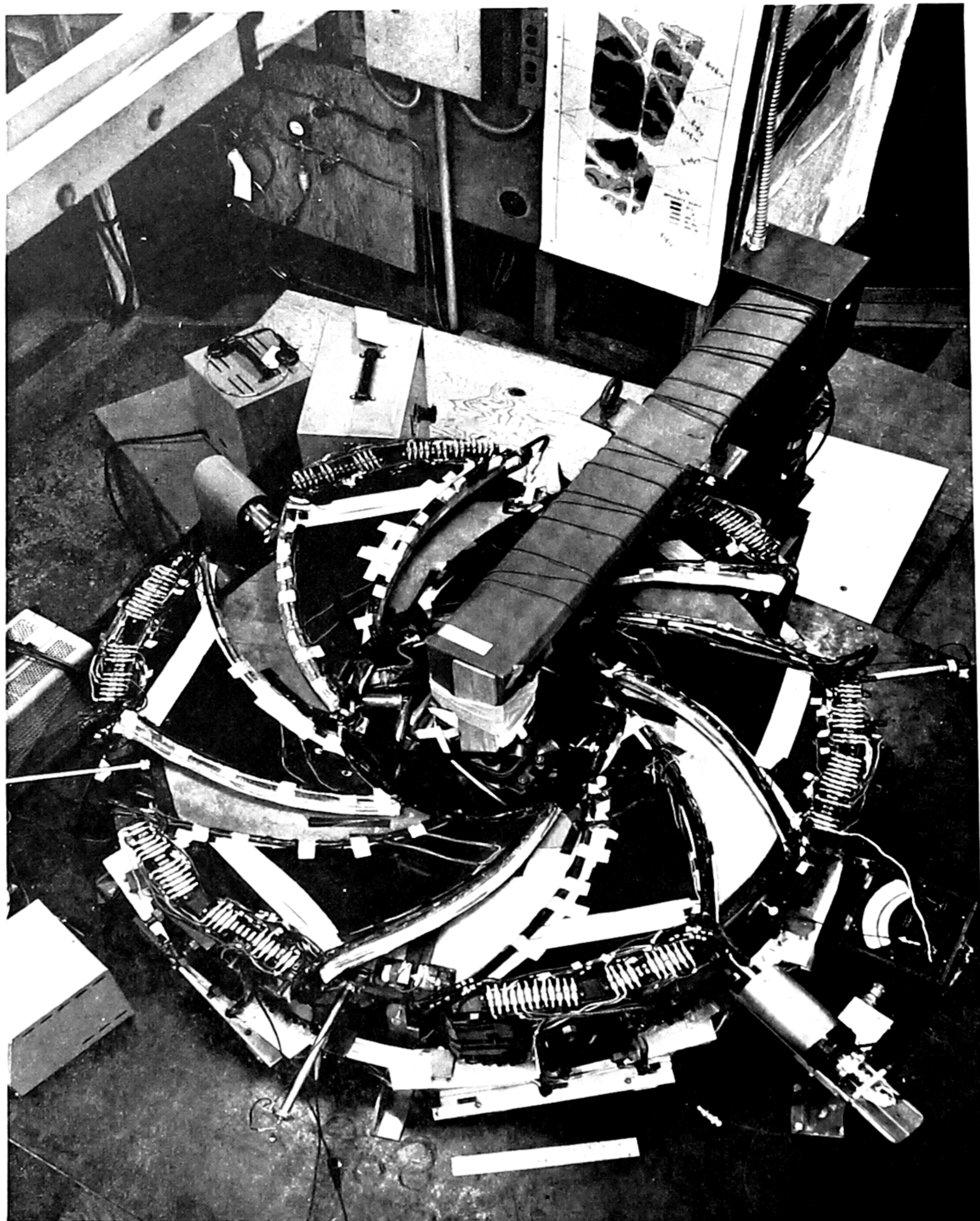
Early in this article I mentioned that two machines now under construction, one at Brookhaven National Laboratory



COSMOTRON, the 3-bev proton synchrotron at Brookhaven National Laboratory, was the first one of the multi-bev accelerators. Its 2,200-ton magnet has an inside diameter of 60 feet.



PHASOTRON is a 10-bev proton synchrotron in the U.S.S.R. Its magnet, of which a portion appears in this photograph, weighs 36,000 tons and is approximately 200 feet in diameter.



FFAG (fixed-field alternating-gradient) design is embodied in an electron accelerator built as a model for a larger proton machine

at the laboratory of the Midwestern Universities Research Association in Madison, Wis. The dark spiral sectors are the magnets.

and the other at CERN in Geneva, will produce protons of 25 to 30 bev. Both of these machines are proton synchrotrons; each will cost between \$20 million and \$30 million. The diameter of the orbit traveled by their protons will be nearly 1,000 feet!

These machines were made possible by the discovery at Brookhaven of a new principle called strong focusing [see "A 100-Billion-Volt Accelerator," by Ernest D. Courant; *SCIENTIFIC AMERICAN*, May, 1953]. This principle involves a reshaping of the guiding magnetic field so that the particles are held much closer to their ideal orbit. It means that the doughnut can be thinner, and the surrounding magnet smaller and lighter.

Until now we have considered only the radius of the orbit, *i.e.*, the size of the circle on which the particles travel. However, the particles can not only drift in and out but also up and down; thus they must be focused vertically as well as horizontally. In old-style synchrotrons the lines of force in the magnetic field are bowed slightly outward [see diagram on page 434]. This has the effect of forcing particles toward the center line when they move above or below it. But the bowed field gets somewhat weaker with the distance from the center line. Hence a particle that wanders too far from the center line is not strongly pushed back toward it.

In strong focusing the field is broken into sectors which are alternately bowed outward and inward [see diagram on page 435]. The sectors bowed outward provide sharp vertical focusing, but are even worse than the old field-shape at bringing a particle in from an orbit that is too large. In other words, they do not focus radially. On the other hand, the sectors bowed inward increase in strength as the radius gets bigger, and provide strong radial focusing. Vertically, however, they have the wrong effect on the particles, tending to spread rather than to focus them. It turns out that each of the defocusing influences is overbalanced by the focusing effect of the other sector; the net result is a much more tightly restricted beam. This method of focusing was successfully used in the Cornell 1-bev electron synchrotron, and it will be applied in the 6-bev Harvard-M.I.T. electron synchrotron.

Not to be outdone by CERN and Brookhaven, the U.S.S.R. has announced that it will build a 50-bev strong-focusing proton synchrotron. The magnet will weigh about 22,000 tons and will have a diameter of 1,500 feet. It would seem that whatever we do, our Soviet friends

can do too—and with a factor of two in their favor.

"FFAG"

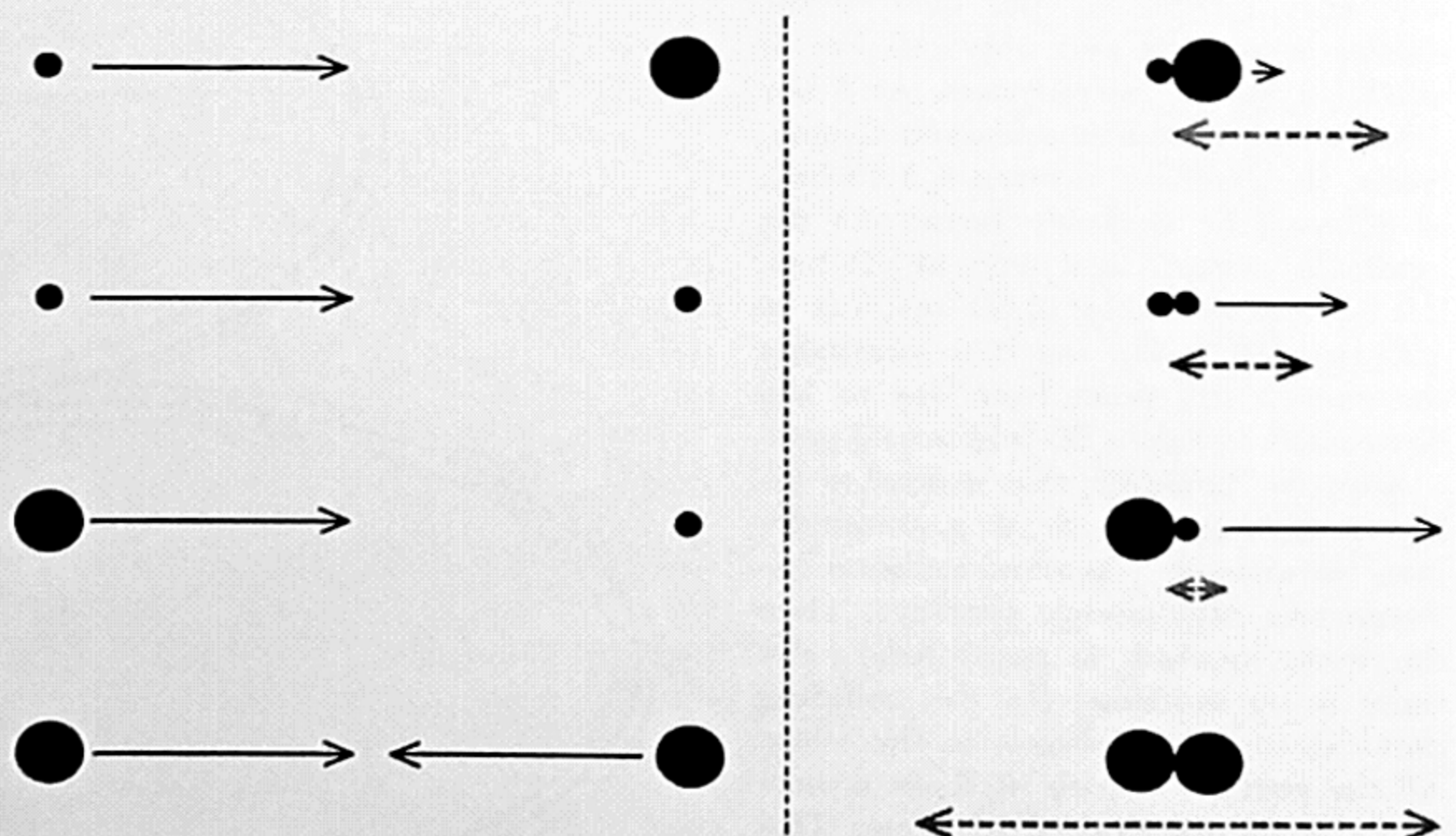
The most exciting recent development in this country has been the "fixed-field alternating-gradient" accelerator proposed by Keith R. Symon of the Midwestern Universities Research Association (MURA). The so-called FFAG machine is really a rococo cyclotron in which the magnetic field is shaped in such a way as to allow the cyclotron to work into the high-energy relativistic region. We have already seen how the ordinary cyclotron is limited to accelerating protons to about 20 mev. When this limitation was first pointed out in 1938, L. H. Thomas of the Ohio State University suggested a way to get around it. He proposed to scallop the pole tips of the cyclotron magnet so that the surfaces would consist of a series of ridges running out from the center, with valleys in between. Thomas showed that the strength of the resulting field would increase toward the outside, compensating for the protons' relativistic increase in mass, and would also focus the protons so that they would stay in the vacuum chamber. Thomas's scheme was far too complicated for the techniques of the time, and it was ignored. Now we realize that he had anticipated the strong-focusing principle. Two Thomas-type cyclotrons are now under construction, one at Oak Ridge National Labora-

tory, the other at Berkeley. Both of them will produce protons and deuterons in the range of several hundred mev.

We can now understand an FFAG type of accelerator if we imagine that the radial scallops of the Thomas magnet are twisted into spiral ribs. (Is this the flamboyant style that presaged the end of the Gothic period?) The twisting introduces an additional kind of strong focusing. In fact, the idea grew out of strong focusing; only later was its similarity to the Thomas cyclotron recognized. The idea of FFAG has been exploited to the full by the workers of the MURA laboratory at Madison, Wis. They have imagined and computed (using two high-speed computing machines) all sorts of variations of the FFAG geometry, and have built several models that have successfully demonstrated the practicality of the scheme.

The advantage of the fixed-field design is twofold. First, it is easier to control a constant field than a varying one. Second, the fixed-field machines can be operated continuously, whereas the synchrotrons and synchro-cyclotrons must operate cyclically, or in pulses, a new cycle starting each time the field reaches its maximum value. Continuous operation means that more accelerated ions are produced per unit time; in other words, the beam has a higher intensity.

According to the MURA workers, the increased intensity that can be obtained with FFAG machines will make it possible to circumvent a serious limitation



USEFUL ENERGY in a collision depends on the motion of the particles after impact. Solid arrows at left represent energy of motion of bombarding particles. Solid arrows at right show energy of motion of the system after impact. Broken arrows indicate fraction of total energy available for desired reactions. Small dots are light particles; large dots, heavy ones. When like particles are made to collide head-on (*bottom*), all of their energy is available.

on accelerators which I have not mentioned as yet. This limitation concerns the amount of energy that is actually available to produce the reactions we are looking for.

When a high-energy ion from an accelerator strikes a stationary target particle, part of the energy goes into moving the target, and is wasted. It is as if we were trying to break a stone by hitting it with a hammer. To the extent that the hammer blow simply moves the stone, the energy is not available for breaking it. Now if the hammer is very light and the stone very heavy, we can see that the target will not move very far; almost all the energy of the hammer will go into breaking or chipping the stone. If we use a heavy sledge on a light pebble, most of the energy goes into moving the stone, and very little of it is available for breaking the stone. If the hammer and stone weigh the same, they will tend to move off together with half the speed of the incoming hammer; exactly half the energy will be available for breaking the stone.

It is the same with atom-smashing. But here relativity plays a particularly dirty trick, robbing us of most of the advantage to be gained by increasing the energy of the bombarding particles. We have seen that really high energies mean an increase in mass. Thus as we go up in energy we increase the weight of our "hammer" and lose a larger and larger fraction of its energy. At 1 bev a proton is already noticeably heavier than when it is at rest; when it hits a stationary proton, 57 per cent of the energy is wasted and only .43 bev is available for useful purposes. At 3 bev (the energy of the Brookhaven Cosmotron), the available portion is 1.15 bev; at 6 bev (the Berkeley Bevatron) the available portion is 2 bev; at 10 bev, 2.9 bev are available; at 50 bev, 7.5; at 100 bev, 10.5. We see that increasing the energy 100 times from one to 100 bev results in only a 20-fold actual gain.

Suppose, however, that instead of firing a moving particle at a stationary one, we arrange a head-on collision between two high-energy particles. Then the mass increase is neutralized, and there is no tendency for the colliding particles to move one way or the other. All the energy of both of them is now available for the desired reactions. This is what the MURA designers propose.

They have envisaged a bold design, called "synchroclash," in which two 15-bev accelerators are placed so that their proton beams intersect and the particles collide with each other. This will yield an available energy of 30 bev, whereas

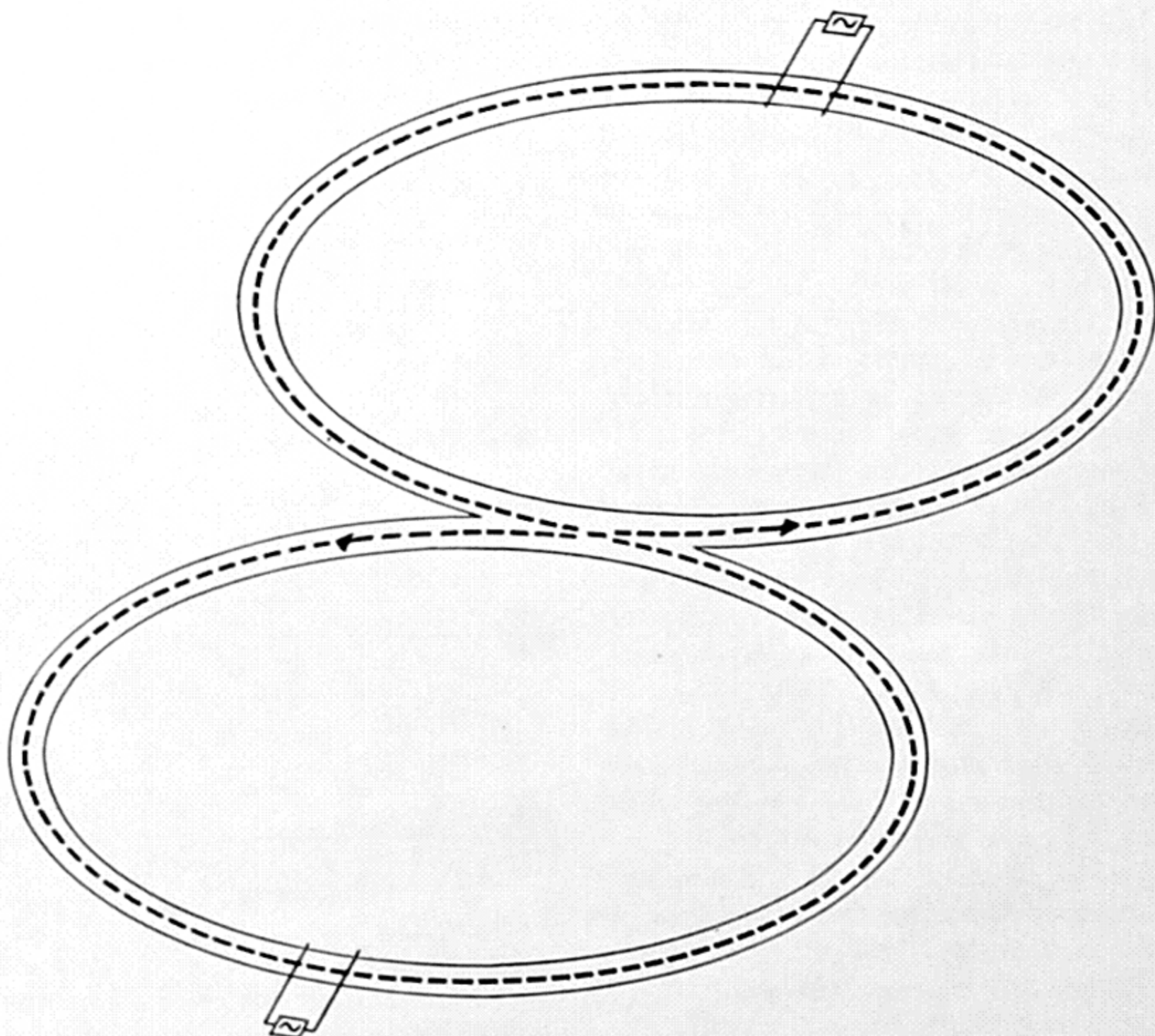
in the case of a 30-bev proton colliding with a proton at rest only 6 bev would be available. In fact, to attain a useful energy of 30 bev in the ordinary way would mean using at least 500 bev. The success of the synchroclash idea turns on the intensity of the accelerator beams: there must be enough protons to make collisions reasonably frequent. The MURA proposal languished for several years, but interest in it seems to have revived. Perhaps the complicated orbits of the artificial satellites have had something to do with the new willingness to consider attempting the complicated orbits of FFAG.

Soviet Ideas

The Soviet designers have gone off in different directions. Veksler has been thinking of a scheme in which one approaches the ideal accelerator, namely one in which the accelerating field appears exactly in the vicinity of the ions but nowhere else. He envisages a small bunch of ions in a plasma (a gas of ions) exciting oscillations or waves in an electron beam. These waves are to act together coherently to give an enormous push to the ions being accelerated. If this is not clear to the reader, it is

because it is not clear to me. The details have managed to escape most of us because of a linguistic ferrous curtain, but Veksler speaks of the theoretical possibility of attaining energies up to 1,000 bev. The proof of the idea must wait until it is put into practice. It should be remarked, however, that other wild schemes of Veksler, for example the synchrotron principle, are incorporated into most of our conventional accelerators today.

G. I. Budker of the U.S.S.R. has also presented some speculative ideas which have obviously been inspired by efforts to produce controlled thermonuclear reactions. Budker proposes an intense circular electron beam maintained by a weak magnetic guide field. The high current of the beam will cause it to "pinch" to a very small diameter because of its own magnetic field. The idea then is to use the very strong local magnetic field around the pinched beam as the guide field of a conventional accelerator [see diagram on page 441]. With an electron beam six meters in diameter one could expect to hold protons with an energy as high as 100 bev. Budker and his colleagues have constructed a special accelerator in which they have achieved a 10-ampere current of 3-mev



SYNCHROCLASH design would set two accelerators side by side so that their beams overlapped. Head-on collisions between particles would provide the maximum of useful energy.

electrons, and they expect to attain much higher currents and energies before long. It could well be that something really revolutionary will come out of this energetic work.

Our own thermonuclear program has inspired research on very strong magnetic fields [see "Strong Magnetic Fields," by Harold P. Furth et al.; *SCIENTIFIC AMERICAN*, February, 1958]. It seems likely that this development will find an application to the guidance of particles in multi-bev accelerators.

Electron Accelerators

These new machines we have been discussing are proton accelerators, but there is vigorous activity in electron machines as well. We have already mentioned the Harvard-M.I.T. synchrotron which will attain 6 to 7.5 bev, and the half-dozen other smaller machines in the billion-volt range. The 220-foot linear electron accelerator at Stanford University has been on the scene for some time. Its energy has steadily increased so that it may now be used in experiments at 600 mev. We expect to welcome it to the 1-bev club before long.

The linear machine is significant because there is a special difficulty in reaching high energy with electron synchrotrons. When electrons are made to travel on a curved path at high speeds they give off strong electromagnetic ra-

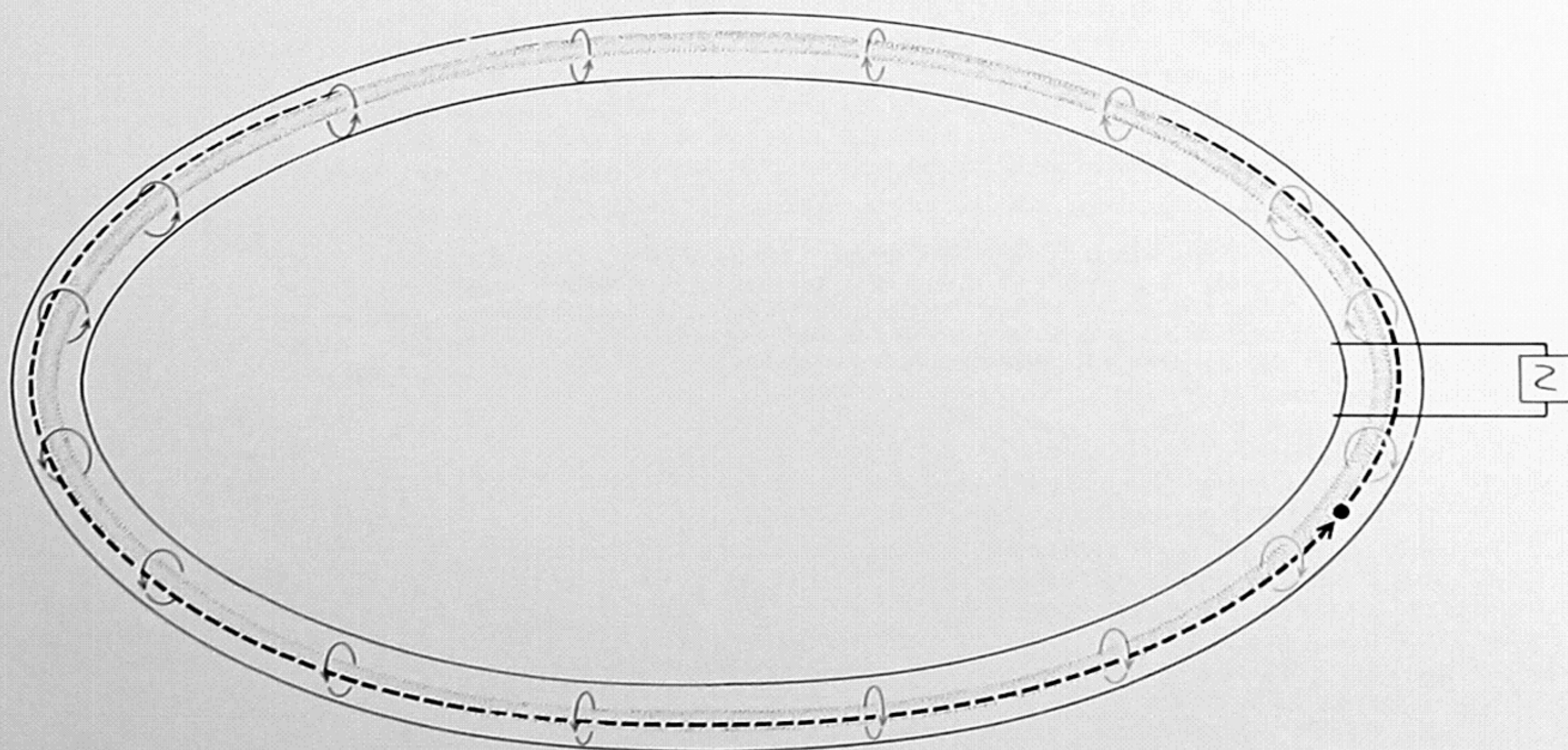
diation. The effect is easily visible to the naked eye. The difficulty is that this radiation can represent a substantial loss of energy, and it increases rapidly as the energy of the machine goes up. In the Harvard-M.I.T. synchrotron the amount of energy radiated is almost prohibitive (about 10 mev per turn at 7.5 bev). To reach higher energies, say 20 bev, the Stanford group has been thinking in terms of a linear accelerator, which does not have this radiation difficulty because its particles do not move in a circle. Such a machine might be as much as three miles long.

I am not convinced that the limit of electron synchrotrons has been reached. Indeed, it is not difficult to imagine a 50-bev electron synchrotron. The radiation problem would be solved by reducing the curvature of the electron beam, that is, by increasing its radius to, say, half a mile. I believe that the upper limit of the electron synchrotron may be as high as 100 mev.

While we are "thinking big" we should not forget Enrico Fermi's proposal to ring the earth with a vacuum tube and, using the earth's magnetic field, obtain 100,000 bev. For that matter, now that artificial satellites are commonplace, we might put up a ring of satellites—each containing focusing magnets, accelerators, injectors and so on—around the earth. Something like a million bev could be expected from this

accelerator, which we might as well call the lunatron. At the very least such a device will eliminate the need for vacuum pumps, since it will be outside the atmosphere.

Villard de Honnecourt and later Viollet-le-Duc have left us detailed accounts of the builders of cathedrals and of their methods. It seems to be pretty much the same story then and now. The designer of the cathedral was not exactly an architect, nor is the designer of an accelerator exactly a physicist. Both jobs require a fusion of science, technology and art. The designers of cathedrals were well acquainted with each other; the homogeneity of their work in different countries is evidence of a considerable interchange of information. The homogeneity of accelerator design demonstrates the same interchange today. Our medieval predecessors were only human; one gets the definite impression that they were subject to petty jealousies, that occasionally there was thievery of ideas, that sometimes their motivation was simply to impress their colleagues or to humiliate their competitors. All these human traits are occasionally displayed by their modern counterparts. But one also gets a strong impression of the excitement of those mighty medieval creators as they exulted in their achievements. This sense of excitement is no less intense among modern nuclear physicists.



PINCH EFFECT might be used to provide a magnetic guiding field for an accelerator, thus eliminating the heavy magnet. The

dotted ring is a pinched plasma. Its magnetic field, which is shown by colored lines, would act to hold particles near its outer edge.

The Author

ROBERT R. WILSON moved so many times in his childhood that his school record lists seven grammar schools and five high schools in various parts of Wyoming, Colorado and California. This resulted in poor grades which made it difficult for him to enter college. Yet before joining the freshman class of the University of California he had already built a rudimentary laboratory in which he made high-vacuum pumps of his own design. Wilson majored in electrical engineering, but a near-failing grade in freshman physics challenged him to study physics instead, against the advice of his teachers and in spite of the handicap of poor health. By his senior year he was doing original research in gaseous discharges under Ernest O. Lawrence, and had discovered an original way to study the time lag of the spark discharge. As a graduate student at Berkeley Wilson turned to nuclear physics and was the first to work out the theory of the cyclotron. In 1940 Wilson joined the faculty of Princeton University. There he invented a way to separate uranium iso-

topes, with the result that at the age of 28 he found himself in charge of a \$500,000 project employing 50 people. Wilson and his staff were soon moved to Los Alamos, where he led the cyclotron group. Eventually he became head of the Experimental Nuclear Physics Division and mayor of the community. Since World War II Wilson has taught at Harvard University and Cornell University. At Cornell he built the first strong-focusing synchrotron and has experimented with thermonuclear reactors, none of which have worked so far. He is one of the founders of the Federation of American Scientists. In his spare time he likes to carve sculptures in wood.

Bibliography

THE CYCLOTRON. W. B. Mann. Methuen & Co., Ltd., 1940.

HIGH-ENERGY ACCELERATORS. M. Stanley Livingston. Interscience Publishers, Inc., 1954.

NUCLEAR REACTORS FOR RESEARCH. Edited by Clifford K. Beck. D. Van Nostrand Company, Inc., 1956.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

THE TRACKS OF NUCLEAR PARTICLES

by Herman Yagoda

Each constituent of the atom is characterized by its trajectory through matter. Here these signatures are discussed with special reference to their visualization in thick photographic emulsions.

A nuclear physicist studying the elementary particles of nature is in much the same position as an explorer trying to picture unknown animals from their tracks. The physicist never can see the particles themselves—only their footprints in a cloud chamber or a photographic plate. But from these tracks he deduces a particle's mass, movements, speed, lifetime and social impact on its fellows. By now the tracks of some members of the nuclear family are almost as familiar and readable as the footprints of a domestic animal. Interesting new tracks keep turning up, some strange, some predictable—the latest to make its appearance is that of the long-sought antiproton. It seems a timely moment to survey the scene and review the gallery of footprints that identify the members of the strange population in the nucleus of the atom.

We shall consider the tracks as they are recorded in photographic emulsions. It was in this medium that the existence of particles in the nucleus of the atom was first detected—through the fact that Henri Becquerel left some uranium near photographic film in a drawer. Becquerel noted simply that radioactive emanations from the uranium had fogged his film. That the "fog" might consist of a network of tracks was not discovered until 13 years later. In 1909 Otto M \ddot{u} gge of Germany exposed some film to tiny crystals of zircon, a feebly radioactive mineral. To study the faintly developed image he had to use a microscope, and he then noticed that there were fine linear tracks radiating from the crystals. Not long afterward the tracks of alpha particles emitted by radium were recorded in fine-grained emulsions at Lord Rutherford's famous laboratory in England.

When a charged particle travels through a photographic emulsion, it

forms a latent image in the silver bromide grains, just as light does. In the case of the moving particle, the latent image results from ionization by the particle along its path. This image, marking the track of the particle, is then made visible by development of the emulsion in the usual way. So that fast particles may be brought to a stop within the emulsion, it is usually made as thick as possible. Emulsions used to track cosmic rays and high-energy particles from accelerators are often more than one millimeter thick—about 100 times thicker than in ordinary photographic film. The length of a particle's track in the emulsion must be measured precisely to determine its kinetic energy. Since the path slants into the emulsion, its length cannot be measured directly: it is computed by means of the Pythagorean theorem from the two measurable distances—the depth at which the particle comes to rest in the emulsion and the horizontal distance along the emulsion surface from the point of entry to the point directly above the end of the track.

At best the search for particle tracks in emulsions is slow, tedious work. It takes many hours or days of poring over the photographic plate with a microscope to find and trace the faint lines of silver grains. For this reason physicists long preferred to use cloud chambers for particle detection work. But the photographic plate has an obvious advantage over a cloud chamber. Particles traveling through this denser medium are more likely to collide with atomic nuclei and produce interesting developments. A great deal of work has been done to improve nuclear emulsions. In 1947 Pierre Demers of the University of Montreal found a way to prepare stable emulsions containing 90 per cent silver bromide, instead of the usual 30 per

cent, and in these more concentrated emulsions particles produce more robust tracks.

Let us proceed to examine some of the identifying tracks. We shall begin by immersing a photographic plate in a very dilute solution of a soluble compound of the radioactive element radium. After leaving it for a time (days, weeks or months) in a dark place, we remove the plate, develop it and inspect it under a microscope. Here and there on the plate we see starlike sets of short heavy tracks, each set radiating like spokes from a hub point. The tracks identify the particles as slow alpha particles, and the formation is known as an alpha star. At the center of the star a radium atom has emitted a series of alpha particles. The radium atom decays first to radon, then to other unstable descendants and finally to lead. In this spontaneous transmutation from radium to lead a total of five alpha particles (plus several beta particles) is emitted. Each in the series comes out with a characteristic kinetic energy, and the different energies (ranging up to 7.7 million electron volts) cause the tracks in a star to be of different lengths.

Occasionally the star seen in a photographic plate may represent the disintegration of not one but many radium atoms. This was made clear by an experiment performed by Mlle. C. Chamie at the Curie Institute in Paris. She exposed a plate in an extremely dilute solution of polonium, the last alpha-emitting descendant of radium in the transition to lead. It was supposed that single tracks of alpha particles, from separate atoms of polonium, would appear in the emulsion. Instead Mlle. Chamie found stars consisting of several hundred alpha tracks from a common center. All the

tracks were of the same length, corresponding to the energy of alpha-emission from polonium. Evidently even in an extremely dilute solution the polonium atoms are not completely dissociated into individual ions but may cluster in groups of several thousand atoms. The collections have been named radiocolloids.

All matter contains traces of radioactive substances, and their energy fields have been pulsating in minerals since the earth's crust solidified eons ago. Nature strews the investigator's path with clues—if we could only see. Long before the discovery of radioactivity, geologists had observed that grains in certain minerals, such as mica, were sometimes surrounded with halos of colored material. They could find no way to explain how these colored bands might be formed. In 1907, when radioactivity was a topic of growing interest, John Joly in Ireland noted that the distance from the center of each tiny sphere to the halo around it was about the same as the range of an alpha particle emitted by radium or thorium. He suggested what is now taken to be the correct solution of the mystery: that alpha particles radiating from radioactive atoms at the center ionize iron atoms in the mica near the end of their path, cause the iron to become oxidized and thereby produce the colored bands.

Just as familiar, and as ubiquitous, as

the footprints of alpha particles are the footprints of beta particles, or electrons. These light particles make very faint, highly scattered tracks in an emulsion. Originating from radioactive substances and from cosmic ray showers, flying electrons record their presence in emulsions wherever placed or however carefully shielded. Even at great depths underground a photographic plate will show about one million electron tracks per cubic centimeter for each day of its underground exposure.

No footprints are more fascinating than those of the strange particles known as mesons. Had present emulsions been in use in the 1920s, their tracks would have been discovered first and "explained" afterward; as it was, the particles were predicted by the theoretician Hideki Yukawa two years before they were actually found. Yukawa invented the meson to account for the binding force that holds particles together in the atomic nucleus. Tracks of a particle such as he had predicted—about 200 times heavier than the electron—were first discovered in 1937 in cloud chambers monitoring the products of cosmic rays. A mystery soon developed: the theory said that these particles should interact strongly with atomic nuclei, but experiments proved that they were rarely absorbed by nuclei.

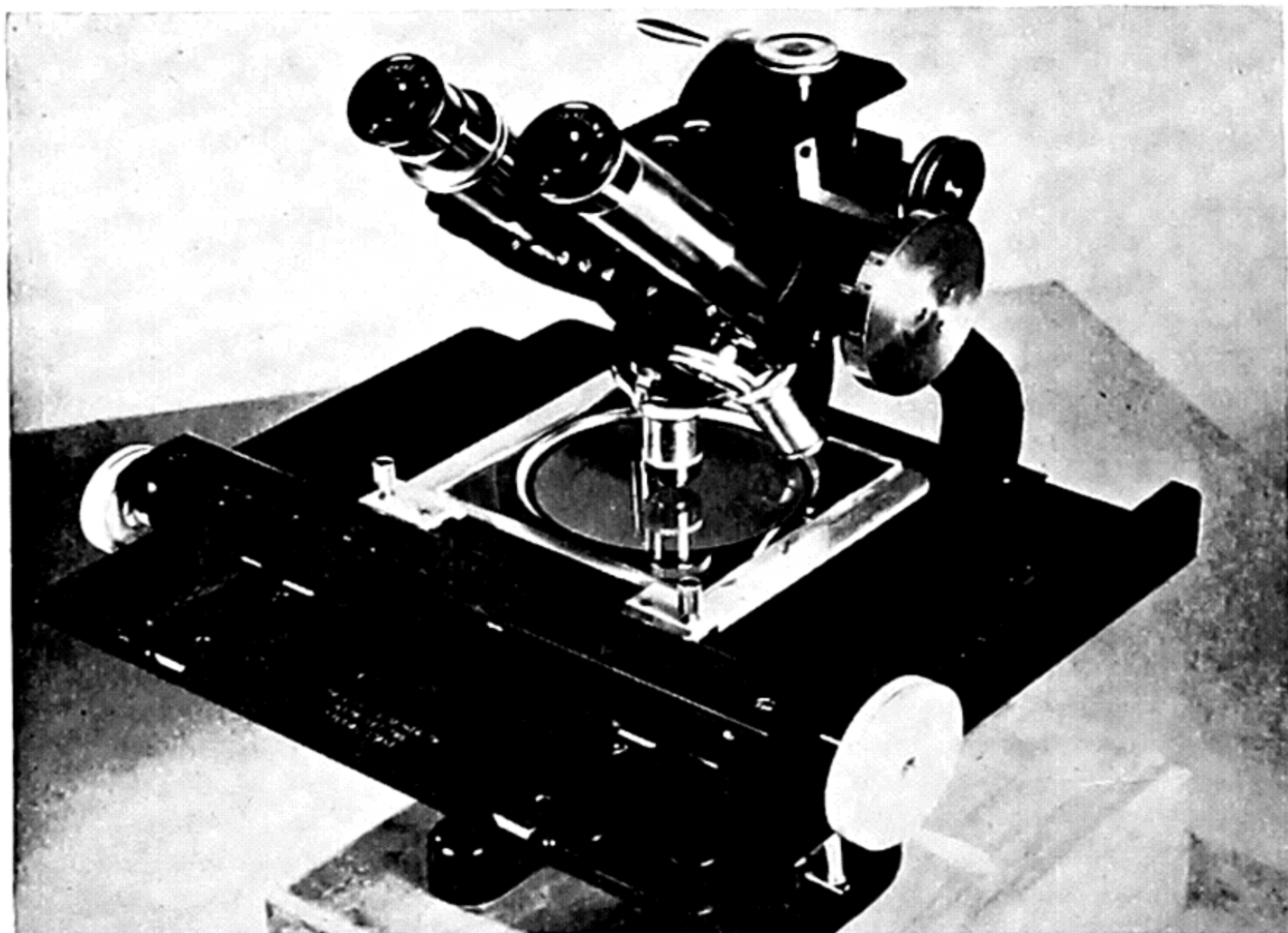
While the theoreticians were pondering this hiatus between theory and experiment, the younger physicists were busy climbing mountains and exposing photographic plates to the intense cosmic radiation high in the atmosphere. By 1947 they had discovered a second, heavier meson which did react strongly with matter [see "The Multiplicity of Particles," by Robert E. Marshak; *SCIENTIFIC AMERICAN*, January, 1952]. A Bristol University team of investigators headed by C. F. Powell obtained photographs showing that when the heavy pi meson came to rest it promptly decayed into the lighter mu meson.

A year later the young Brazilian C. M. G. Lattes, a member of the Bristol cosmic ray group, came to the University of California and in cooperation with Eugene Gardner succeeded in detecting mesons from nuclei attacked by a 400-million-electron-volt beam of alpha particles from the Berkeley cyclotron. Two types of pi meson tracks were then identified. Positively charged pi mesons decayed into mu mesons. Negatively charged pi mesons reacted with atomic nuclei, and the disintegration of the capturing nucleus produced a star.

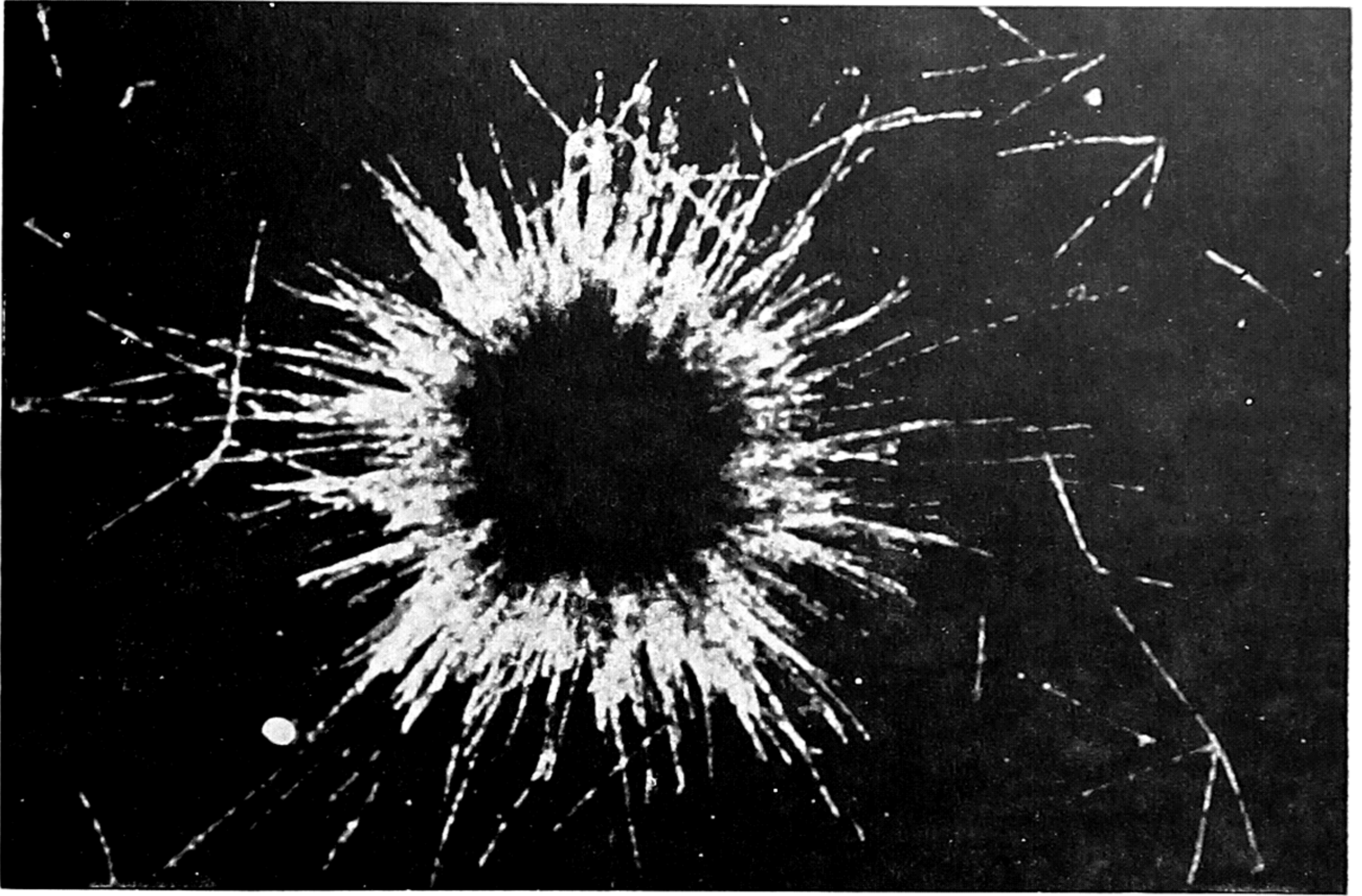
Meanwhile the European investigators, lacking funds for the construction of expensive accelerators, continued to study mesons in the cosmic radiation—the poor man's cyclotron. These simple experiments gave birth to a perplexing number of new particles.

Their first addition to the growing fraternity of Greek-lettered mesons was the tau particle. The Bristol University investigators found its track in an electron-sensitive plate exposed beneath a 12-inch-thick block of lead at the Jungfraujoch High Altitude Research Station. The particle, heavier than a pi meson, produced an unusual three-pronged star on coming to rest. All three prongs could be identified as the tracks of pi mesons. From the available evidence Powell came to the conclusion that the tau meson was an unstable, singly charged particle about 1,000 times heavier than the electron. Powell's brilliant deductions tempt one to finish off his description with the admiring exclamation: "A new particle—elementary, my dear Watson!"

The heavy tau meson is very rare, but an extensive vigil has now detected a number of these events and established the particle's properties. Recent controlled experiments with the six-billion-electron-volt Bevatron at Berkeley indicate that the tau particle and certain other heavy mesons (known as K mes-



SPECIAL MICROSCOPE is used to scan nuclear emulsions. The large stage enables the viewer to follow long tracks. Here the emulsion is a disk embedded in a rectangular Lucite frame fitted with a cover glass. The depth of the track is read on the wheel at upper right.



ALPHA PARTICLES made the image in this dark-field photomicrograph. The emulsion itself contains tiny colloid particles of radi-

um, one of which is at the center of the image. The tracks were made by alpha particles emitted by radium and its daughter elements.



ALPHA STARS emerged from thorium atoms in this emulsion. The stars at left and right represent the serial decay of single thor-

ium atoms. First the thorium atom emitted an alpha particle, then the daughter isotope emitted another alpha particle, and so on.

ons) probably are all the same particle showing alternate modes of decay.

Neutral particles unfortunately leave no footprints in an emulsion or cloud chamber. They may, however, signal their presence indirectly. For example, a fast neutron charging through an

emulsion may collide head on with a hydrogen atom, rip away the latter's electron and cause the proton to recoil and make a track that tells the story of the collision.

At Berkeley all eyes are focused just now on the footprints of the antiproton, which at long last was generated by the

Bevatron a few months ago. The anti-proton—the negatively charged counterpart of the positive proton—has only a fleeting life, but it makes its existence unmistakably known by the spectacular manner of its death. When the particle comes to rest in an emulsion, there is an explosion which generates a large star.

GROUP	MEMBERS	SYMBOL	REST MASS (ELECTRON MASSES)	MEAN LIFE (SECONDS)
NUCLEONS	PROTON	p^+	1836.13	STABLE
	ANTIPROTON	p^-	1840 ± 90	$\sim 5 \times 10^{-8}$
	NEUTRON	n^0	1838.65	750
LEPTONS	ELECTRON	e^-	1	STABLE
	POSITRON	e^+	1	ANNIHILATES
	NEUTRINO	ν	0	
LIGHT MESONS	NEGATIVE PI MESON	π^-	272.8 ± 0.3	2.44×10^{-8}
	POSITIVE PI MESON	π^+	273.3 ± 0.2	2.53×10^{-8}
	NEUTRAL PI MESON	π^0	263.7 ± 0.7	5×10^{-15}
	NEGATIVE MU MESON	μ^-	207 ± 0.5	
	POSITIVE MU MESON	μ^+	206.9 ± 0.4	2.15×10^{-6}
HEAVY MESONS	TAU MESON	τ^+	965.5 ± 0.7	$\sim 5 \times 10^{-8}$
	THETA MESON	θ^0	965 ± 10	1.6×10^{-10}
	CHI MESON	$\chi (K\pi_2)$	963 ± 9	1×10^{-8}
		$(K\mu_2)$	960 ± 7	1×10^{-8}
	KAPPA MESON	$K (K\mu_3)$	955 ± 9	1×10^{-8}
		(Ke_3)	~ 960	
HYPERONS	LAMBDA PARTICLE	Λ^0	2182 ± 2	3.7×10^{-10}
	POSITIVE SIGMA PARTICLE	Σ^+	2327 ± 4	$\sim 10^{-10}$
	NEGATIVE SIGMA PARTICLE	Σ^-	2325	$\sim 10^{-10}$
	CASCADE PARTICLE	Ξ^-	2582 ± 10	$\sim 10^{-10}$

FUNDAMENTAL PARTICLES are listed, together with their characteristic tracks in nuclear emulsions. The photon and graviton are omitted to simplify the organization of the chart. The light

mesons are called L particles; the heavy mesons, K particles; the hyperons, Y particles. The chi and kappa mesons have dual symbols, the second of which segregates them according to their mode

The particles emerging from the explosion, among which are several pi mesons, have a large kinetic energy; the total energy released is about that predicted by the theory that the antiproton and a proton combine and annihilate each other, converting mass into energy.

The Bevatron produces antiprotons

when a beam of high-energy protons (at 6.2 billion electron volts) hits a copper target. The fast protons attacking the nuclei of the copper atoms generate large numbers of heavy mesons and an occasional antiproton: the yield is about one antiproton per 62,000 mesons. The theory suggests that a high-energy pro-

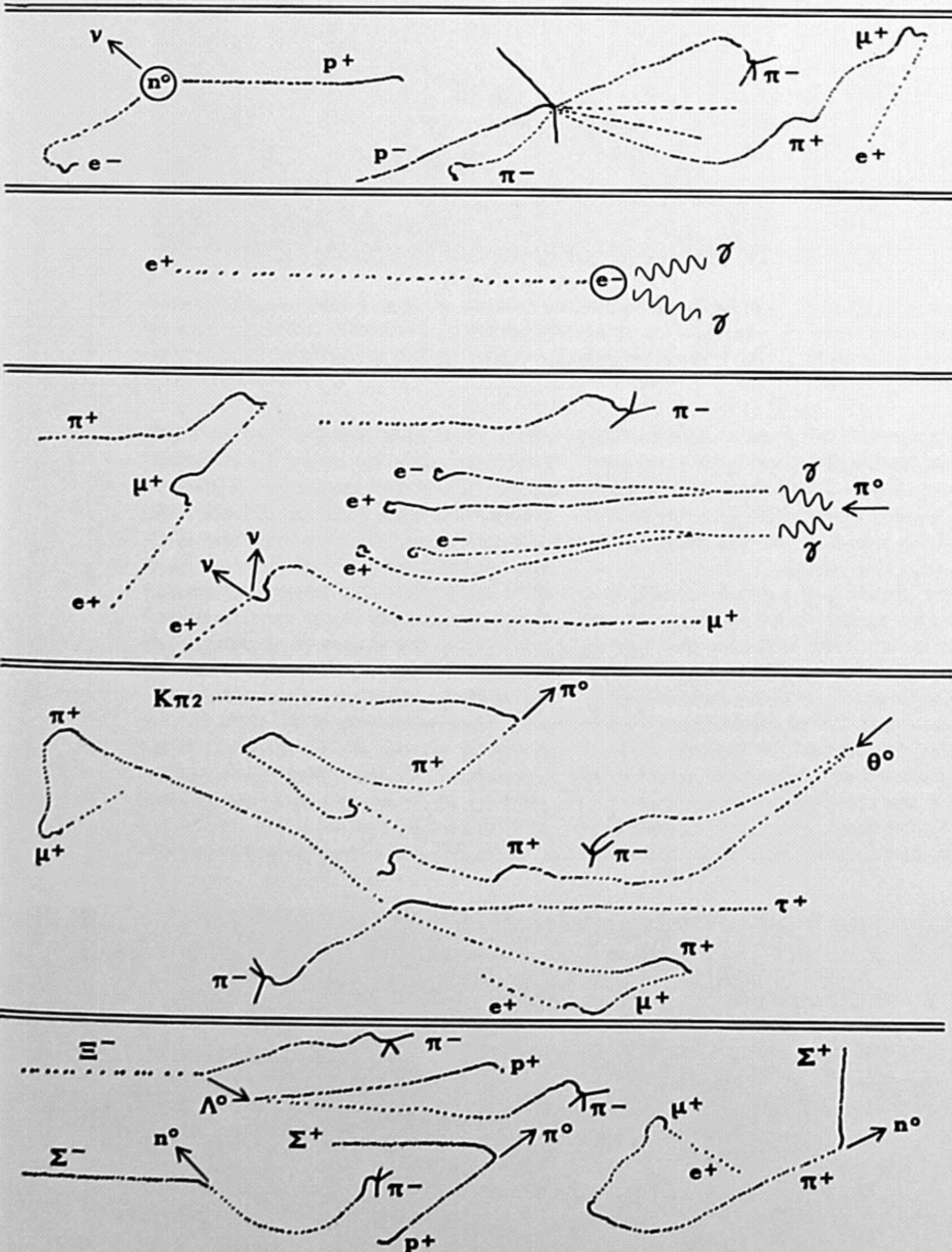
ton interacts with a neutron to form an antiproton-proton pair.

The antiproton has the same mass as a proton. One would therefore expect that it should have about the same probability of collision with atomic nuclei as it travels through matter. But experiments with the new particle show that the antiproton actually has about twice as great a collision probability, or cross section, as the proton. This surprising property has presented nuclear physicists with an intriguing problem.

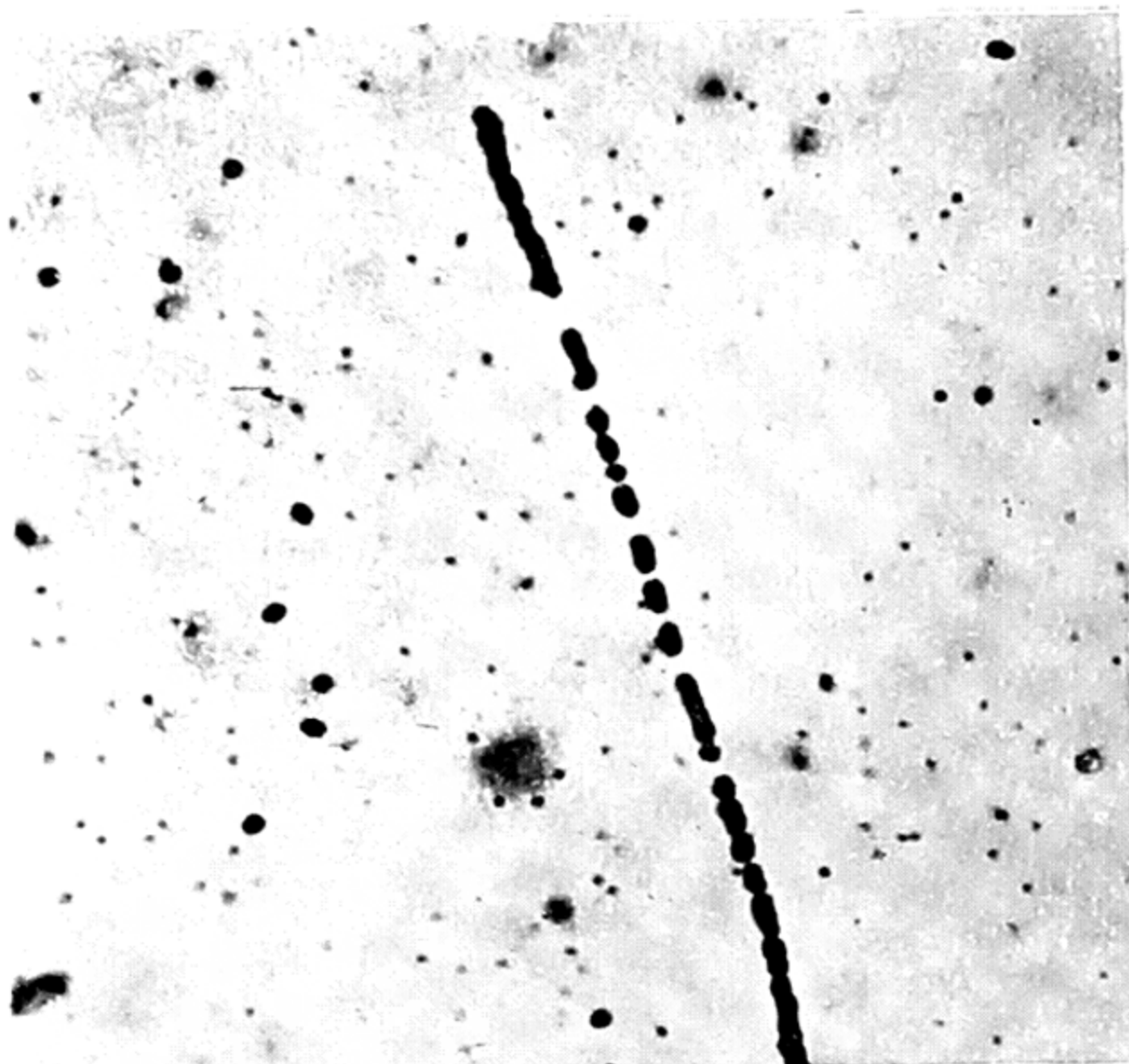
Enlightening as the work with atom-smashing machines has been, the investigators of particles have not by any means lost interest in the wild assortment of nuclei and nuclear debris that rains into our atmosphere from the bombardment of the cosmic radiation. Of the primary cosmic radiation itself, little reaches ground level, for the atmosphere absorbs it as effectively as would a three-foot-thick layer of lead completely surrounding the earth. But physicists are capturing the footprints of primary particles coming in from space by floating their instruments and photographic plates to the top of the air ocean in balloons. Great impetus was given to this work by the U. S. Navy's development of the plastic "Skyhook" balloon. Unlike rubber balloons, the plastic vehicles can be held at a fixed, preset elevation. Stacks of emulsions have been flown to 100,000 feet—almost at the borders of empty space, for the weight of the overlying air there is only 13 grams per square centimeter, as against 1,030 grams at sea level.

As the primary cosmic rays smash nitrogen and oxygen atoms in the air, they generate a fallout of secondary and tertiary particles. The footprints of these fragments are being recorded at mountaintop stations all over the world. Men who risk their lives to climb a mountain simply "because it is there" are usually very cooperative with the cosmic ray physicists. A light package of photographic plates does not add appreciably to the burden of the climb, and it may add incentive as a form of applied mountaineering. In the ascent of Mt. Everest Sir Edmund Hillary took a small package of plates (given him by Professor Eugster of Zurich University) to the 25,850-foot camp site. Unfortunately, in the excitement of the triumphant descent from the peak the plates were overlooked. Sir John Hunt, the leader of the expedition, apologized in his book, *The Conquest of Everest*: "I very much

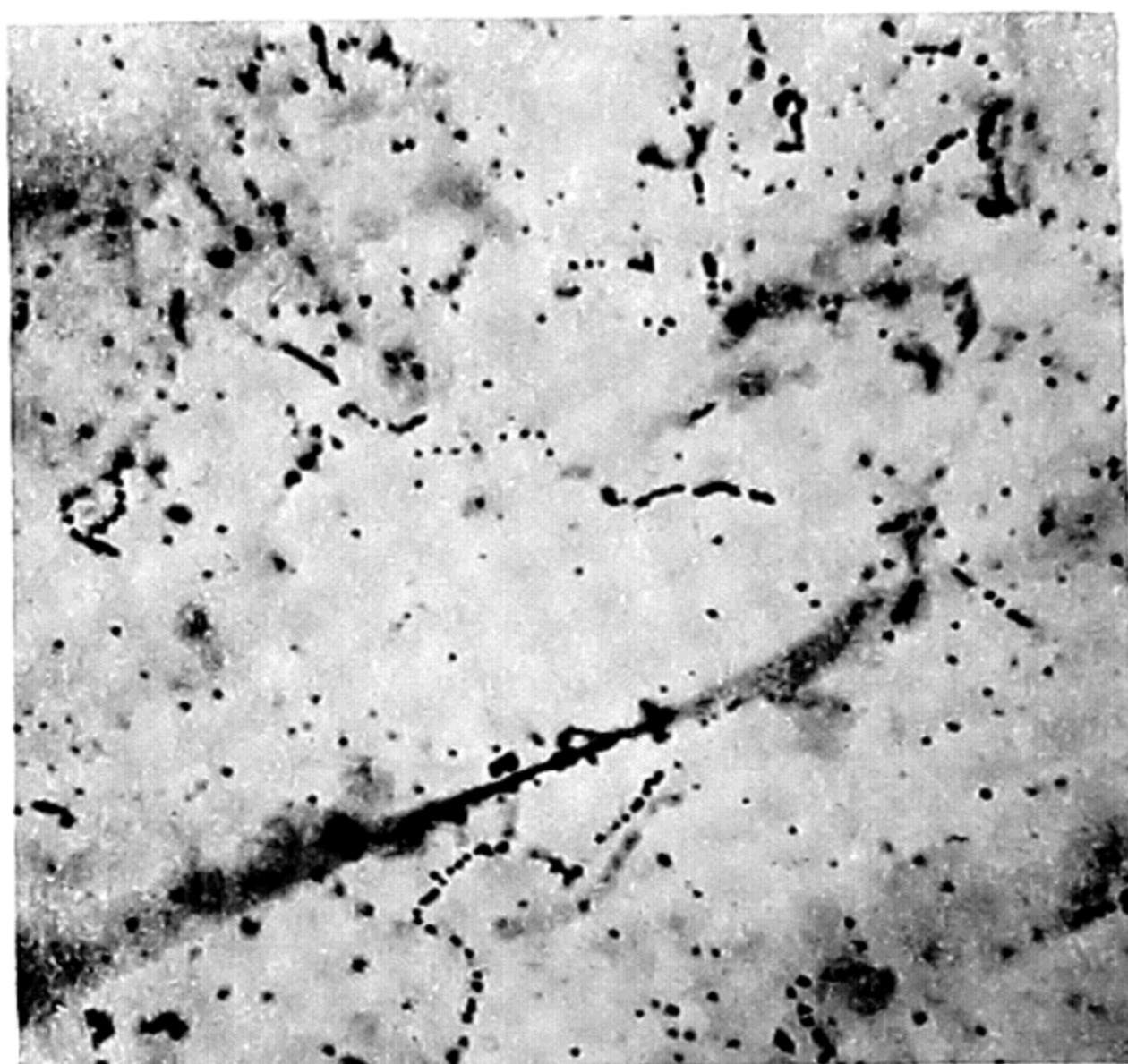
DECAY SCHEME



of decay. $K\pi 2$, for example, indicates that this K (not kappa) particle decays into two pi mesons. The decay schemes may be followed by beginning with the particle in that group. The wavy lines (gamma rays), circles and arrows denote particles that do not make tracks.



SLOW NEUTRON gave rise to this track in an emulsion containing lithium borate. The neutron encountered a lithium atom at the lower end of the short, heavy line at the top. The track was then made by two fragments of the nucleus recoiling from each other.



ELECTRONS made the faint, wavy tracks in this emulsion, which was aged for 50 days before it was developed. The heavy track at the bottom was made by an oxygen nucleus in primary cosmic radiation. The electron tracks along this image are called delta rays.

regret to say that the plates have remained on the South Col, where they must by now have made a very definite recording of . . . cosmic ray phenomena."

Among the first to get a recording of these phenomena was Marietta Blau of the University of Vienna. Nineteen years ago she exposed a series of photographic plates for four months on a mountaintop at Innsbruck. When she developed them, she found not only the familiar alpha stars from radioactive substances but also a number of bigger stars with much longer, less dense prongs. The tracks evidently were produced chiefly by protons. Dr. Blau sur-

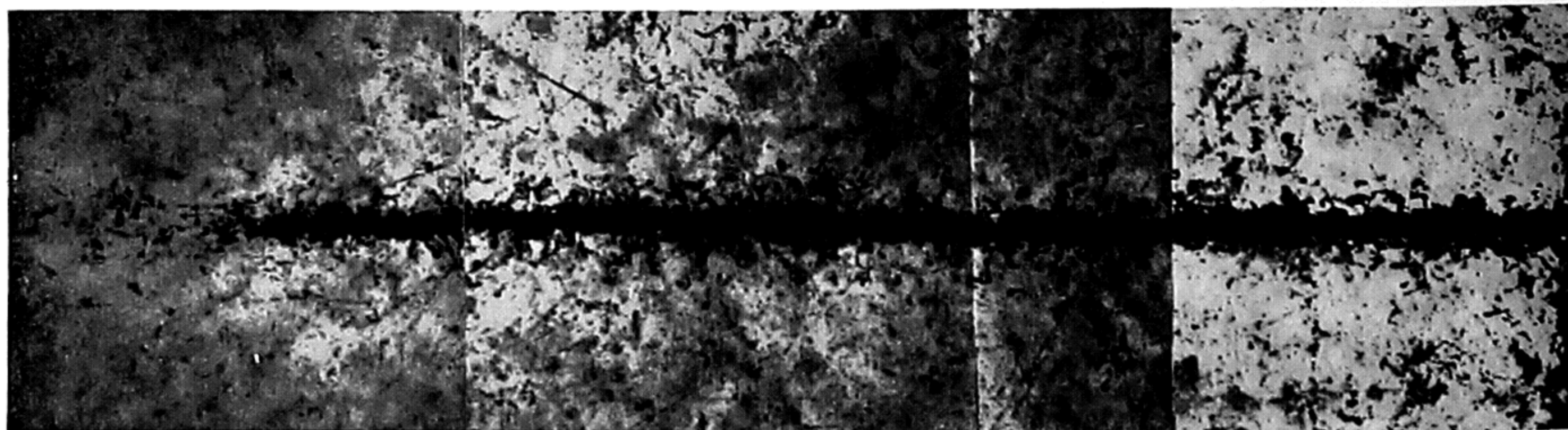
mised correctly that they were the debris of nuclei disrupted by cosmic rays; she followed up this finding and today is studying nuclear disruptions produced by the Cosmotron at the Brookhaven National Laboratory.

The smashing of nuclei by cosmic rays increases rapidly with altitude. At sea level in northern latitudes the rate of star production in photographic plates is about one per cubic centimeter of emulsion per day of exposure; at 14,260 feet on Mt. Evans in Colorado the rate is 20 times that; and in balloons near the top of the atmosphere, 2,500 times.

The tracks of the primary cosmic particles that arrive there from space are

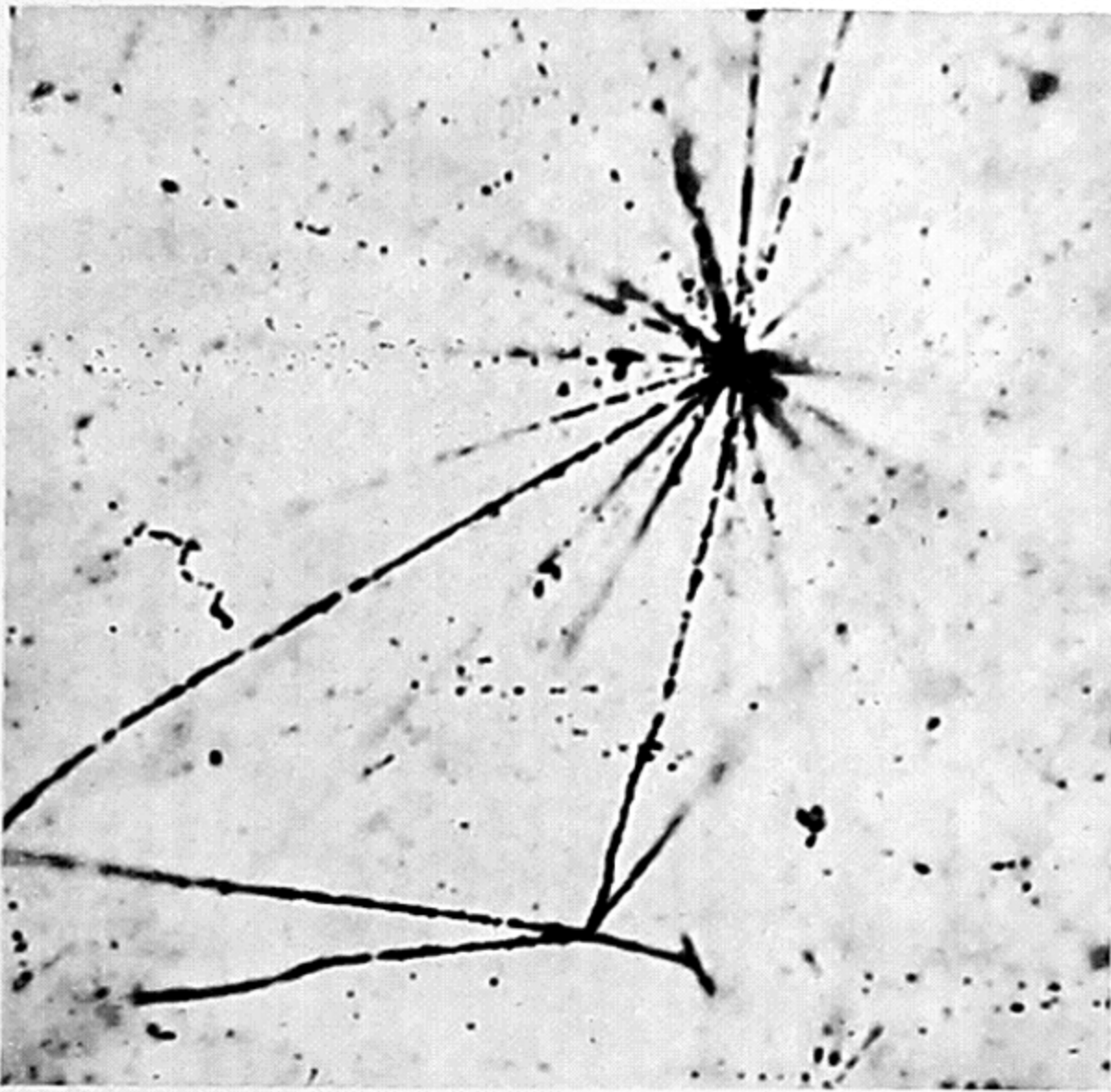
often extremely robust. These thick tracks are made by heavy nuclei, much larger than the nuclei of hydrogen atoms. The track is covered with a fur of spurs projecting from its sides—secondary ionizations which are known as delta rays. Since the amount of ionization by a particle along its path is proportional to the square of its charge, the amount of delta-ray ionization identifies the particle. The primary cosmic particles have been found to include the nuclei of almost all the elements from hydrogen to nickel. Iron nuclei often produce tracks heavy enough to be seen with the naked eye.

Sometimes the incoming heavy nu-

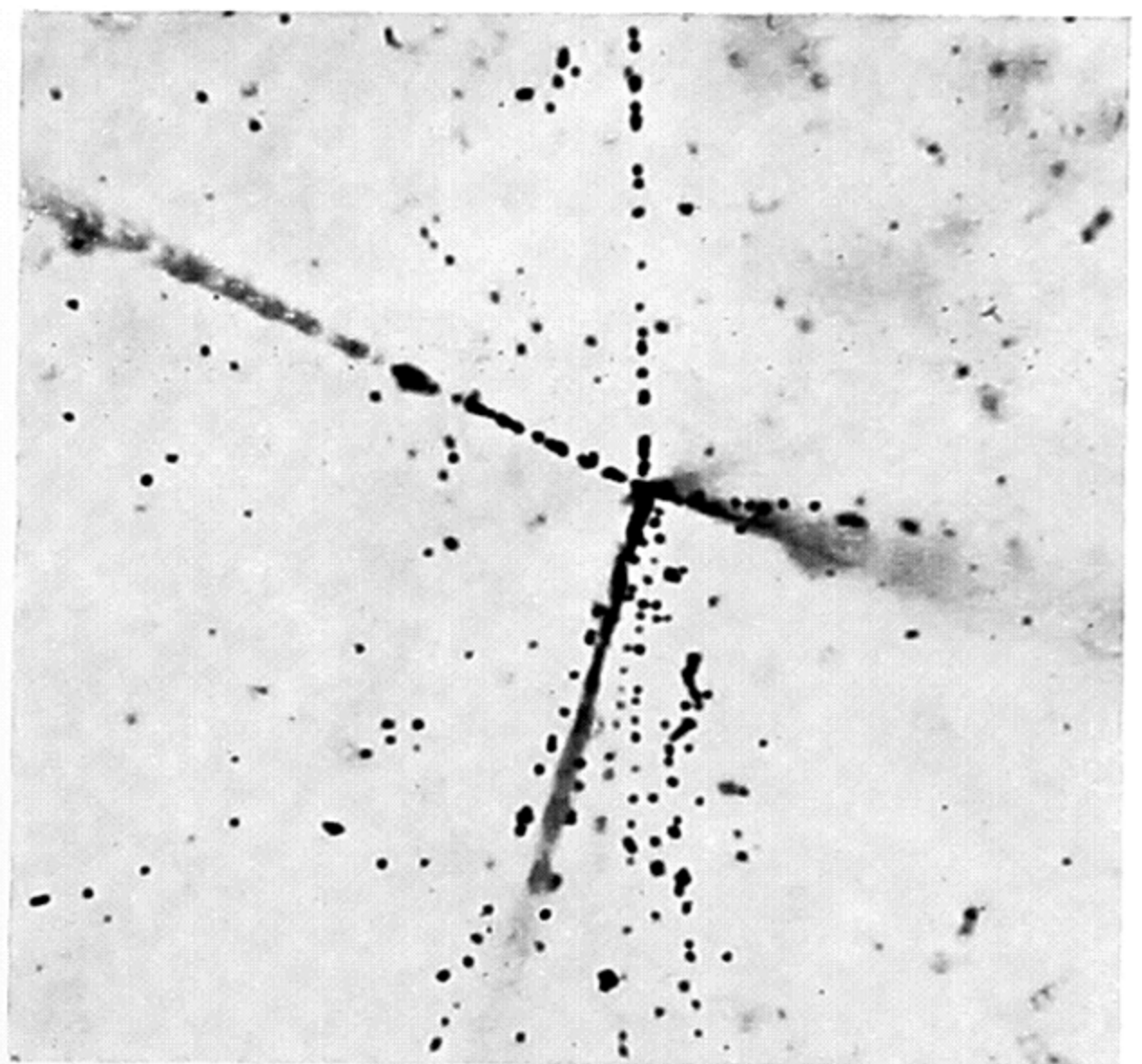


IRON NUCLEUS in primary cosmic radiation entered this picture from the left. Escaping catastrophic collision with nuclei in

the emulsion, it finally came to rest at the right. Its energy was dissipated by a series of encounters in which it removed electrons



NEGATIVE PI MESON made the track between these two stars. At the top is a nucleus disrupted by a primary cosmic ray. At the bottom is a second nucleus disrupted by the pi meson. Negative mesons are readily absorbed by nuclei because of their opposite charge.



PROTON in primary cosmic radiation made the nearly vertical track at the top of this emulsion. The tracks produced by its encounter with a nucleus in the center of the emulsion are characteristic of fragments and/or particles with a single electric charge.

cleus is partly sheared off by a glancing collision in the air, and the separated bundles of nucleons diverge from the point of collision. Sometimes the cosmic primary hits an atom head on and disintegrates it, emitting a shower of heavy mesons: as many as 200 charged mesons have been seen in a single star. Many of the pi mesons decay during flight into mu mesons; the latter, nearly immune to capture by atoms, zip through the atmosphere and often plunge deep into the earth.

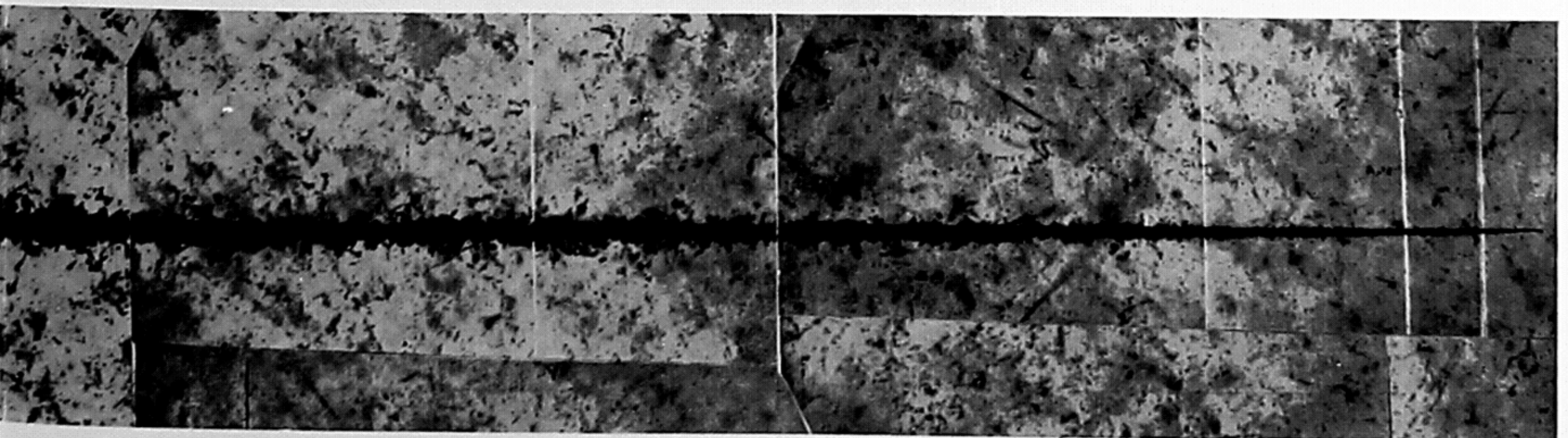
A small proportion of the heavy nuclei from space escape catastrophic collisions and are eventually slowed down by ionization processes in the atmos-

phere. When these particles are caught in an emulsion, they produce very spectacular tracks. The track is first thick and furry; then as the heavy nucleus slows down and begins to pick up electrons, the reduction of its positive charge diminishes the ionization it produces, so that its track tapers down to a needle point at the end of its flight.

The last grain at the rest point of a heavy primary cosmic particle is a thing to marvel at. Embedded within the grain of silver in the emulsion is an atom with a history unlike that of its neighbors. It is an atom which may have been blown out of a star in our galaxy

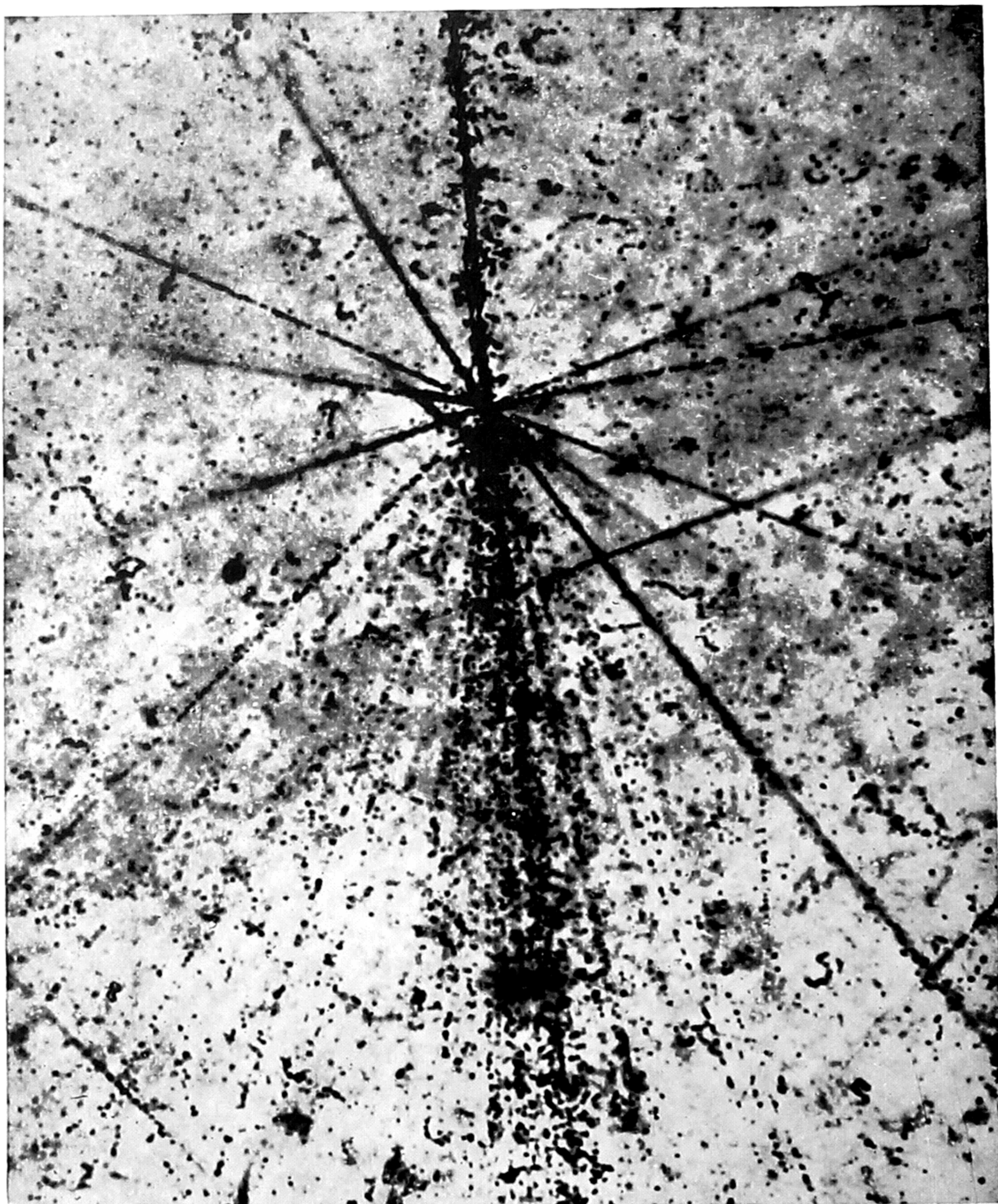
millions of years ago. It was accelerated through interstellar space by magnetohydrodynamic fields. For millions of years it escaped collision with cosmic dust. Finally it plowed into the earth's atmosphere, and in a single moment lost its store of energy accumulated since birth. Such is the ever-increasing entropy of the universe, of which Swinburne wrote:

*We thank with brief thanksgiving
Whatever gods may be
That no man lives forever,
That dead men rise up never;
That even the weariest river
Winds somewhere safe to sea.*



from atoms in the emulsion. These electrons made the wavy tracks along the path of the iron nucleus. The track is about a 16th of an

inch in length, too long to be shown in a single photomicrograph. It has accordingly been depicted in a mosaic of photomicrographs.



GIANT SHOWER OF MESONS is recorded in this photomicrograph of a small section of nuclear emulsion carried to a height of 106,000 feet by a Navy "Skyhook" balloon. At the top of the photomicrograph is the heavy track of an enormously energetic iron nu-

cleus in the primary cosmic radiation. Above the nucleus is a "star" resulting from the collision of the iron nucleus and a nucleus in the emulsion. Below the star is a jet of about 40 pi mesons. To the left and right of the star are heavier fragments of the target nucleus.

The Author

HERMAN YAGODA is a physical chemist at the National Institutes of Health in Bethesda, Md. A New Yorker, he graduated from Cooper Union in 1929 and took his master's degree at New York University in 1931. Later he had a Baker research fellowship at Columbia University. He worked for a chemical company and as an assistant chemist in the U.S. Customs Laboratory in New York before joining the National Institutes of Health in 1942. His interest in cosmic rays and nuclear emulsions grew out of work in microchemical analysis. Yagoda has written a comprehensive book entitled *Radioactive Measurements with Nuclear Emulsions*. He still considers himself a chemist:

"I tell my friends that I have achieved the ultimate goal of microchemists—the ability to identify a single atom of matter."

Bibliography

NUCLEAR PHYSICS IN PHOTOGRAPHS. C. F. Powell and G. P. S. Occhialini. Oxford University Press, 1947.

OBSERVATIONS ON STARS AND HEAVY PRIMARIES RECORDED IN EMULSIONS FLOWN IN VIKING ROCKET NO. 9. Herman Yagoda in *Canadian Journal of Physics*, Vol. 34, No. 1, pages 122–146; January, 1956.

RADIOACTIVE MEASUREMENTS WITH NUCLEAR EMULSIONS. Herman Yagoda. John Wiley & Sons, Inc., 1949.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

THE AGE OF THE ELEMENTS IN THE SOLAR SYSTEM

by John H. Reynolds

Studies of the inert gases found in meteorites have confirmed estimates that the earth is 4.6 billion years old and provide evidence that the elements in the solar system are not much older.

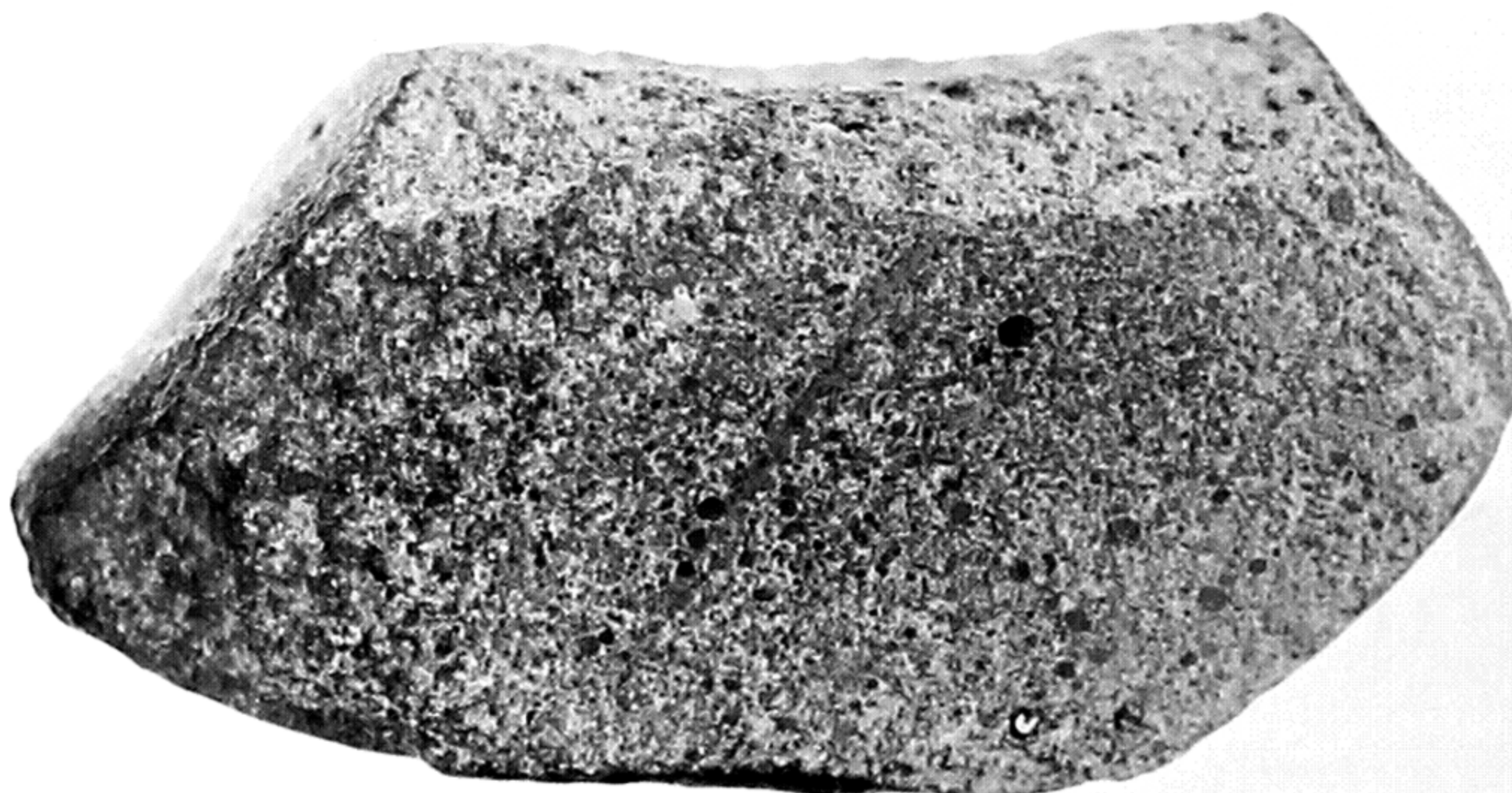
How old is the solar system? The question is a big one, and it suggests research tools of commensurate size—large telescopes on mountaintops and rockets roaring into space. On the contrary, the answer is coming from the analysis of minute bits of matter by quite modest instruments mounted on the laboratory bench. It is true that the bits of matter come from meteorites: those samples of interplanetary rubble that are swept up by the earth's gravitational field. But the passage of cosmic time is measured by counting the few atoms of the "noble" gases—helium, argon, neon, krypton and xenon—that

are trapped in the crystal lattices of meteoritic stone and iron.

Certain isotopes (atoms of the same element that have slightly different masses) of the noble gases represent clocks that have stopped. They are dead ends to which nuclear transformations have carried other atoms higher up in the table of elements. The transformation of some atoms proceeds spontaneously by radioactive decay. Since the rate of decay is immutable and is known, time can be measured by comparing the relative abundance of the parent elements and their noble-gas daughters present in a sample of matter. Some nu-

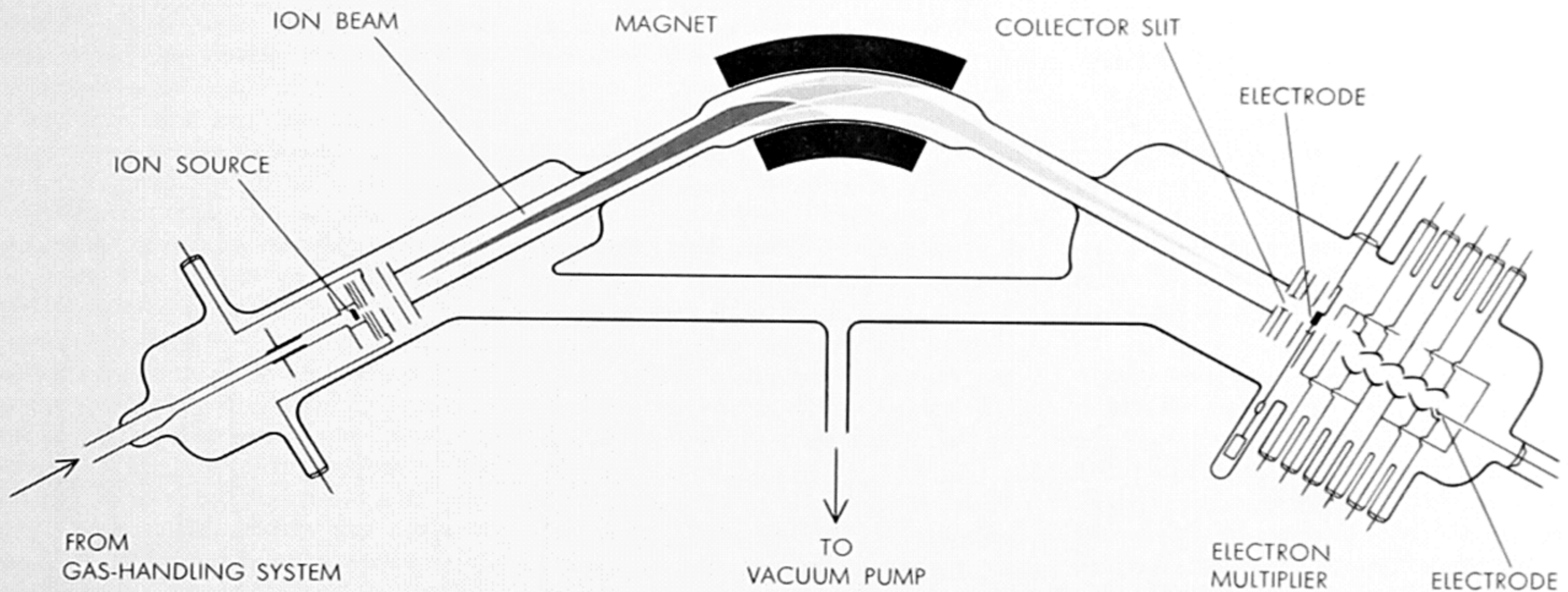
clear transformations in nature are induced by the impact of the highly energetic particles called cosmic rays. The rate at which these latter transformations occur can be estimated within reasonable ranges of error to yield another sort of time-scale.

As investigators have mastered these ways to tell time, they have found in the noble gases an independent check on the ticking of the classic uranium clock in the rocks of the earth. As a result it now appears that the cold planetary bodies of the solar system all crystallized at about the same time: some 4.5 or 4.6 billion years ago. The noble



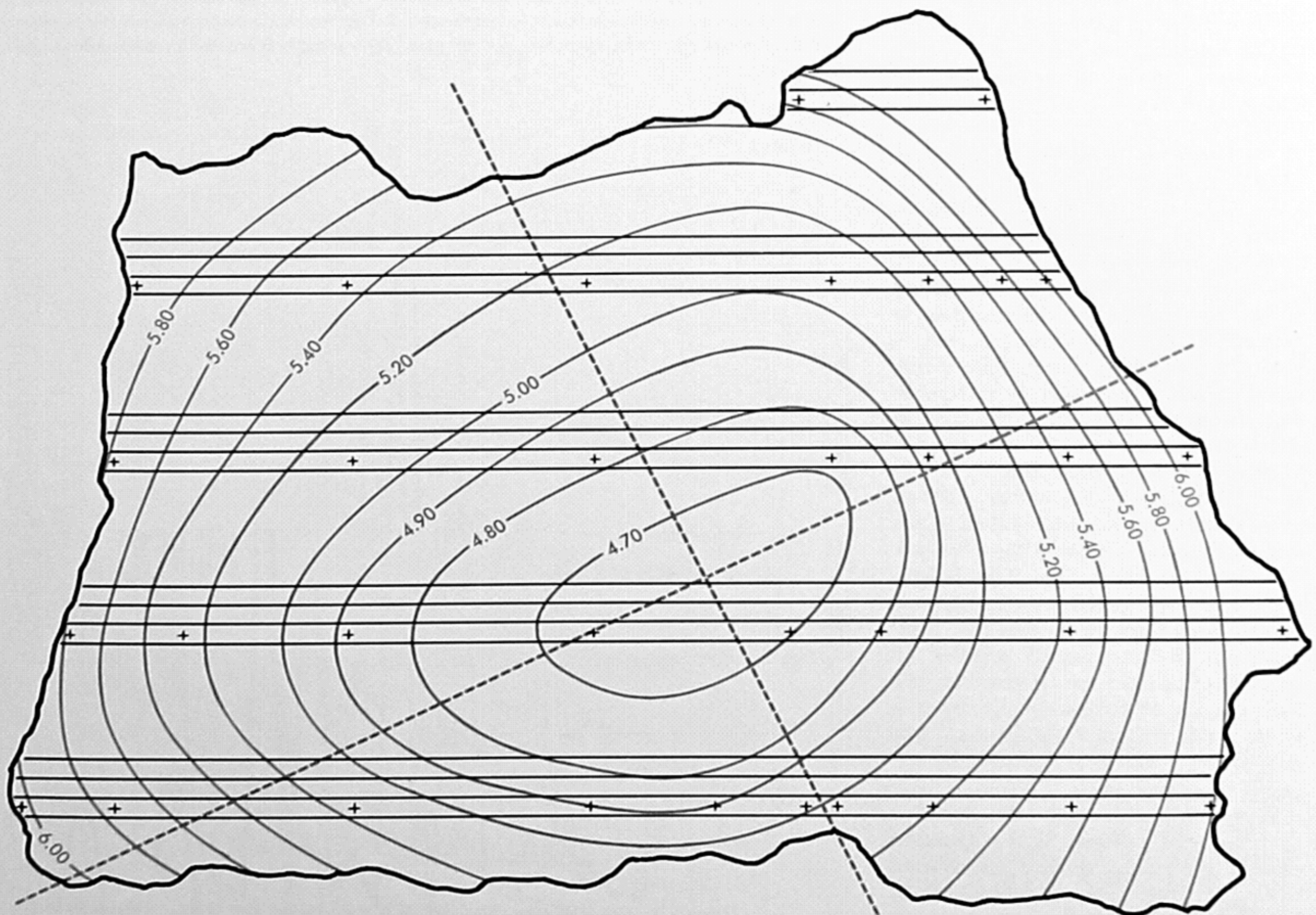
RICHARDTON STONE METEORITE discussed in the text fell in North Dakota in 1918. Specimen shown here, part of 200 pounds of material collected, weighs about 1.4 pounds. Bright area at lower

right is iron inclusion. Smaller sample of the Richardton stone was used by the author in his research on age of elements in the solar system. Photograph was made at the Smithsonian Institution.



MASS SPECTROMETER analyzes small samples of noble gases by distinguishing among gas isotopes according to their weight, or mass. After a virtually permanent vacuum is achieved by pumping and baking, vacuum valve is closed and sample is admitted from gas-handling system. The gas atoms are ionized by a stream of electrons at ion source and accelerated by a series of electrodes, emerging in a diverging stream (*ion beam*). Ions of a particular isotopic

weight are deflected by magnetic field to the collector slit. Focusing action is result of wedge shape of field. Lighter and heavier isotopes are deflected to a greater and lesser degree. Slow variation in the strength of the field causes other isotopes to reach the collector slit in succession, and a mass spectrum is recorded. The ion currents are recorded either at the first electrode or, after intensification by an electron multiplier, at the second electrode.



CONTOURS OF HELIUM-3 CONTENT (solid colored lines) in a slab cut from center of Grant iron meteorite show that present shape is not greatly different from original one. Contours, roughly lines of constant depth in original body, are based on mass-spectroscopic measurements of helium-3 content at points (crosses)

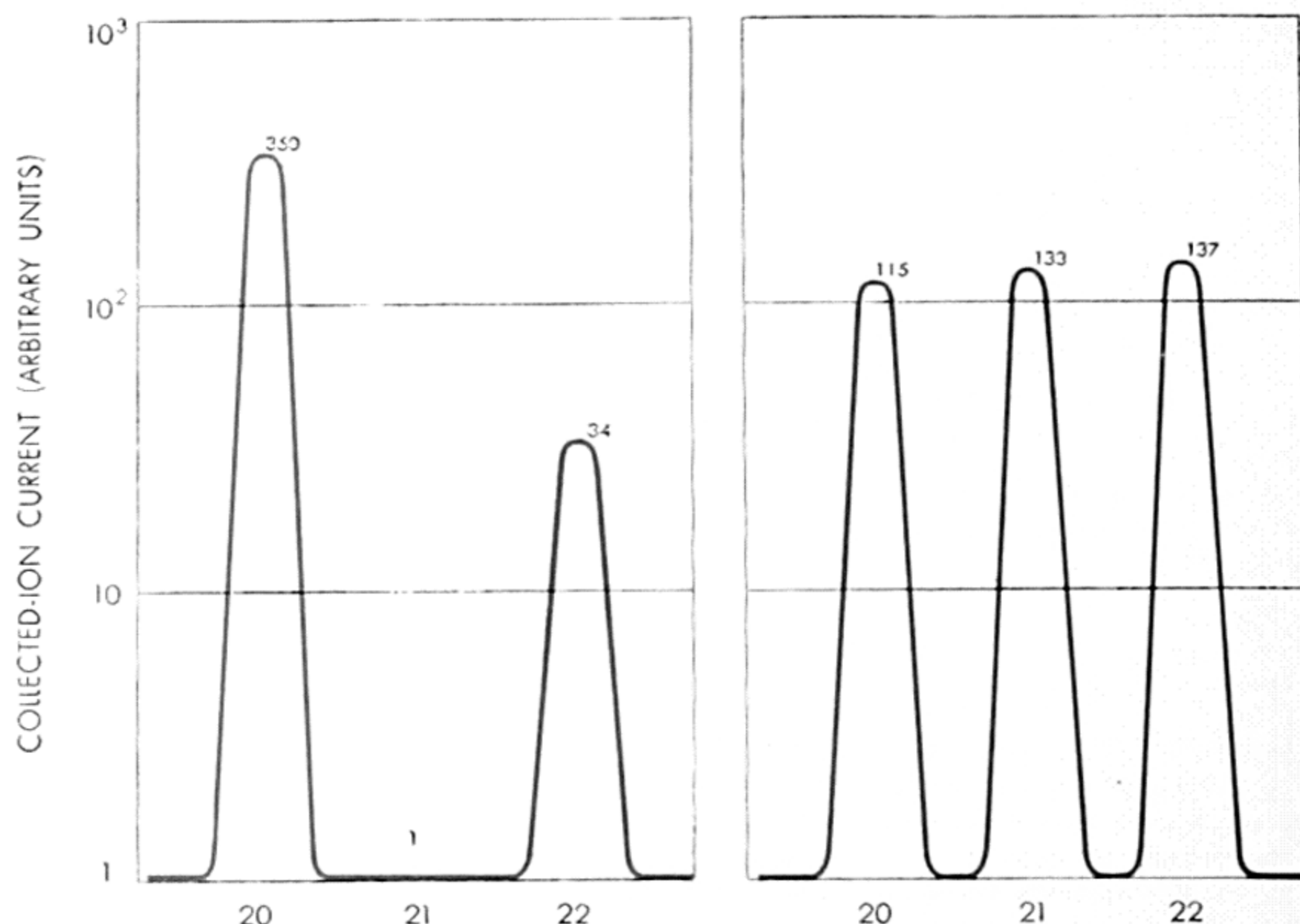
in bars cut from slab. J. H. Hoffman and Alfred O. C. Nier of the University of Minnesota made measurements. Figures give content in millionths of a cubic centimeter per gram. Reduction of content with increased depth is result of increased shielding from cosmic rays that originally produced helium 3 in meteorite.

gas technique also dates the more recent breakup of some of these massive bodies into meteorites and provides clues to the life histories of the fragments. And one noble-gas clock that ran down and stopped at the very dawn of time has made it possible to reach into the twilight of cosmology and measure the full age of the solar system, from the epoch in which the elements that compose it were formed.

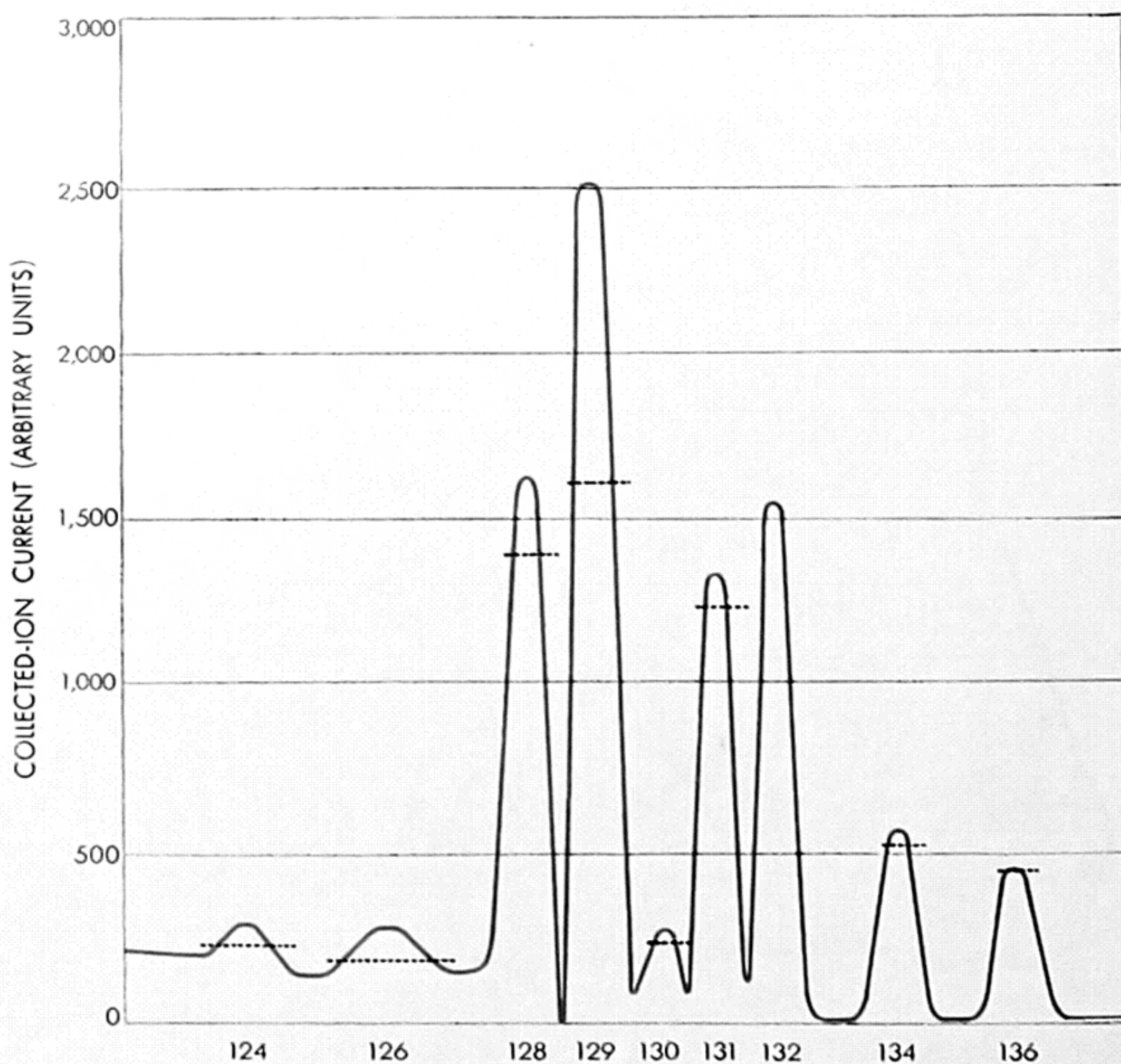
The uranium clock was one of the first fruits of the discovery of radioactivity. As early as 1907 workers in the Cavendish Laboratory of the University of Cambridge recognized that the decay of uranium to helium offered a way to measure the age of rocks. Natural uranium consists of two different isotopes that decay to helium at different rates. The helium-to-uranium ratio in a rock yields, therefore, a measurement of time, although the ratio must be qualified by the realization that the helium tends to leak away. Fortunately the decay of each uranium isotope also yields a different isotope of lead. These lead-to-uranium and lead-to-lead ratios lie behind most of the readings of the uranium clock.

The interest of investigators in helium was renewed about a decade ago, when measurement of the helium-to-uranium ratio in meteorites produced a startling result. In studying the ages of a number of iron meteorites, F. A. Paneth of the University of Durham arrived at values ranging from about one million years to 7.6 billion years. The upper figure raised a paradox; it was more than twice the then estimated age of the solar system [see "The Origin of Meteorites," by S. Fred Singer; *SCIENTIFIC AMERICAN*, November, 1954]. Since there was too much helium in the meteorites with respect to the uranium, it was suggested that some of the helium might have evolved from the breakdown of iron atoms under cosmic ray bombardment as the meteorite traveled on its orbit around the sun. If this were the case, the excess helium would turn out to be partly helium 3 rather than only the helium 4 which terminates the decay of uranium.

However, measuring the relative abundance of isotopes of the elements calls for procedures quite different from those that distinguish one element from another. Since isotopes differ primarily in mass, they must be discriminated by the tools of physics rather than of chemistry. With the help of the mass spectrometer, an instrument that sorts out atoms according to their masses [see illus-



MASS SPECTRUM OF NEON in stone meteorite (*right*) reveals cosmic-ray production of neon isotopes (*numbers at bottom*) in nearly equal abundance. On logarithmic scale used here, neon 21 does not appear in spectrum of equivalent neon sample from earth's atmosphere (*left*).



EXCESS XENON 129 was first found in Richardton meteorite. Here scale of mass spectrum is linear. Xenon 124, 126 and 128 (*numbers at bottom*) are recorded at sensitivity 10 times that for others. Standard was xenon 132. In normal sample peaks would be at dotted lines.

tration on this page], investigators soon established that helium 3 accounted for the paradoxical "age" of the meteorites. This work also had the larger consequence of stimulating studies of the other noble gases that occur in meteorites.

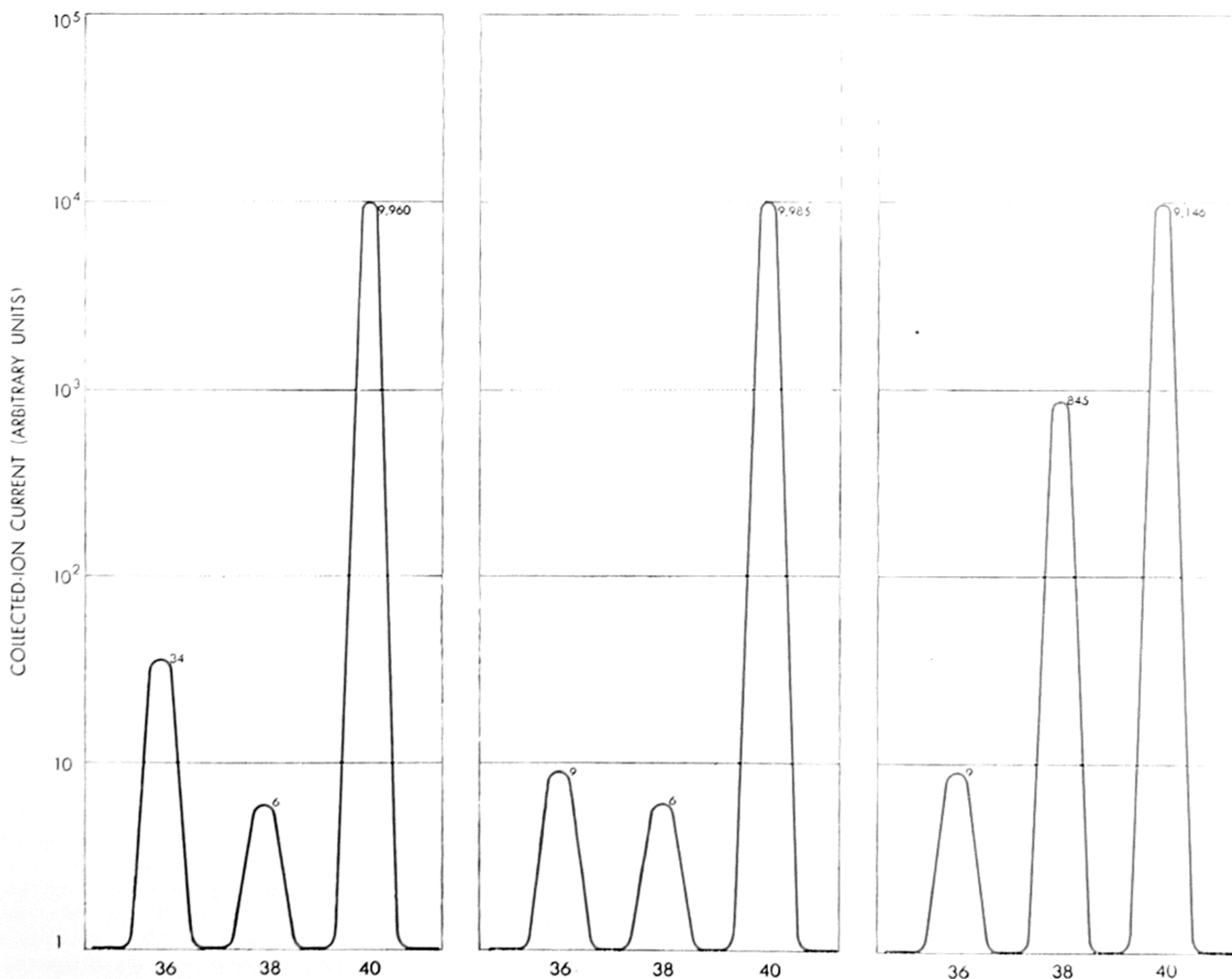
The radioactive isotopes that are produced by cosmic ray bombardment—for example, the carbon 14 in the earth's atmosphere—are easily detected. They proclaim their presence to sensitive radiation counters of various kinds. But when such isotopes are stable, it is normally all but impossible to detect a change in their relative abundance. The change is so small that even over geologic time it lies within the range of error of the best mass spectrometers. Happily the stable isotopes of helium

and the other noble gases in meteorites represent an exception. They are so scarce to begin with that any change in their abundance ratios looms quite large. What is more, the chemical inertness that gives them their patent of nobility makes it possible to isolate the tiny quantities that occur in meteorites, without significant contamination by noble gases of terrestrial origin.

The chemical segregation of chemically active elements by conventional procedures necessarily involves great contamination. Suppose that an element of atomic weight 40 is to be isolated from a sample weighing one gram. The first step is to dissolve the sample in some way, for example by using hydrofluoric acid. The next step will involve

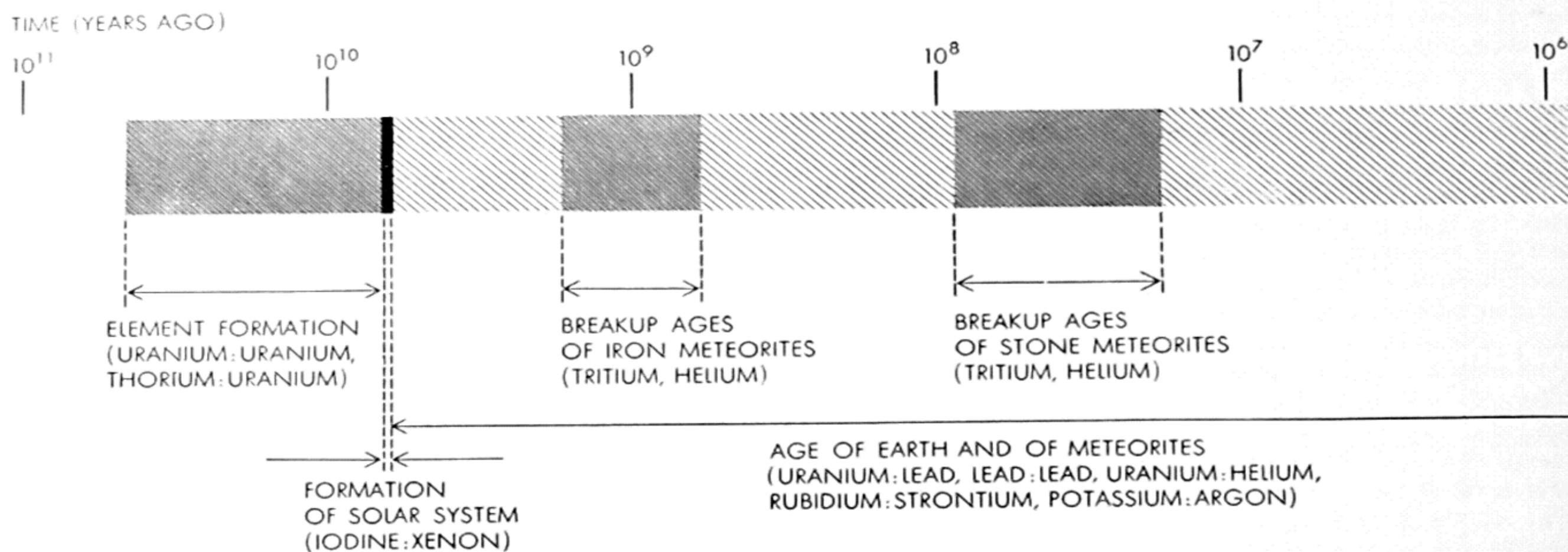
distillation, precipitation or passing the sample through an ion-exchange column. In any case the sample will be mixed with as much as 100 grams of liquid, including a number of extraneous atoms of atomic weight 40. Assuming that a purity (with respect to that atomic weight) of a few parts in a billion can be achieved, the procedure will still introduce about two billionths of 100 grams, or .0000002 gram, of contaminating material. This corresponds to about three million billion (3×10^{15}) atoms of an element of atomic weight 40.

In noble gas "chemistry" the procedure is quite different. To start with, the sample of meteorite is melted in a vacuum. Exposure of the released gas to hot copper oxide, to a trap cooled with



DETERMINING AGE OF METEORITE by one method involves measuring the amount of argon 40 it contains. Argon isotopes in three mass spectra depicted here are identified by mass number at bottom. Two lighter isotopes in spectrum of argon from atmosphere (*left*) are primordial (formed along with other elements). Most of the argon 40, however, has accumulated from radioactive decay of potassium 40. In spectrum of equivalent

amount of argon from a stone meteorite (*middle*) the two lighter isotopes are partially produced by cosmic-ray bombardment of calcium, accounting for the difference in their ratio from that in the first spectrum. In the spectrum at right a known quantity of argon 38 has been added to provide a yardstick for measuring the argon-40 content. The age of the meteorite is calculated from the ratio of this isotope to the potassium-40 content of the meteorite.



HISTORY OF THE SOLAR SYSTEM as calculated from element ratios shown in parentheses is plotted logarithmically. On theory that formation of heavier elements took 10 billion years, interval (*formation of solar system*) from end of element formation to crystallization of earth's crust is 120 million years. On theory that the elements were created at about the same time some five billion years ago, the interval is 290 million years. Age of earth and of

meteorites is 4.6 billion years. Breakup ages of meteorites represent time since parent bodies of which they are fragments last broke up. Lowest value shown for iron meteorites (600 million years) is average breakup age; highest age is 1.7 billion years. Corresponding values for stone meteorites are 20 million and 90 million years. Ages are computed from the content of helium 3 in a meteorite and its rate of production (calculated from tritium decay-rate).

liquid air and to a few milligrams of hot calcium or titanium accomplishes the required purification. All but the noble-gas elements either freeze out in the liquid-air trap or combine with the calcium or titanium, leaving a phase of almost pure noble gas ready for study in the mass spectrometer. Under typical conditions the background pressure in the purification system is about one billionth of an atmosphere. The volume of the system is typically one liter; at this pressure there is about one millionth of a cubic centimeter of extraneous gas. Assuming that the gas has the composition of air, 1 per cent of it will be argon, mostly argon 40. The contamination level at atomic weight 40 is then one hundred-millionth of a standard cubic centimeter, or 300 billion (3×10^{11}) atoms; that is, one ten-thousandth of the contamination in conventional chemistry. For the less abundant noble gases the number of contaminating atoms is correspondingly less. For neon, krypton and xenon the approximate levels of contamination are respectively 500 million, 30 million and two million atoms. Helium is rather a special case because it diffuses so readily in hot solids; special precautions must be taken if helium contamination from the atmosphere is to be kept at a minimum. The last step, measuring the isotopic abundances in the tiny noble gas samples, is accomplished with mass spectrometers especially designed for this task.

Dramatic proof of cosmic-ray pro-

duction of Helium 3 in meteorites came in 1952, when K. I. Mayne of the University of Oxford and Paneth and P. Reasbeck at Durham announced that they had found a helium-3 to helium-4 ratio of about one to three in a number of meteorites. In the earth's atmosphere, by contrast, the ratio is one to a million. Two years later Mayne and Reasbeck found that the ratios among the neon isotopes in meteorites also differ markedly from those in the atmosphere. The neon isotopes of mass 20, 21 and 22 appear in the atmosphere in the ratios of 350 to 1 to 34. In meteorites these three isotopes occur in roughly equal abundance. At about the same time W. Gentner and Josef Zähringer in Germany discovered similar variations in meteoritic argon.

These findings not only brought the ages of meteorites into line with the age of the earth's crust as measured by the uranium clock; they also opened up the history of the meteorites to study. It became possible to determine with considerable accuracy how long a meteorite had been exposed to cosmic radiation before it reached the earth. This involves measuring the amount of helium 3 in a meteoritic sample and dividing this figure by a calculated cosmic-ray production rate. The exposure ages of meteorites vary considerably. In most common stone meteorites it is a few tens of millions of years, with individual ages ranging all the way from four million to 90 million years. The average ex-

posure age of iron meteorites is considerably longer—about 600 million years, with individual ages ranging up to 1.7 billion years.

Though these extreme variations present some difficulties, it is generally agreed that the exposure ages represent breakup ages, that is, the time since a given fragment was last detached from a larger parent body. In a planet or planetoid, the main mass of the material is shielded from cosmic radiation. When such a body breaks up in collisions with other sizable objects, its various fragments are exposed to cosmic rays and, with each successive breakup, still smaller fragments become exposed to bombardment. The systematic difference in the breakup ages of stone and iron meteorites is explained by the deduction that the latter do not break up as easily.

These studies have also told something about the individual life histories of meteorites. By analyzing the distribution of helium 3 in thin cross-sectional slabs of metal cut from two iron meteorites—the Grant meteorite and the Carbo meteorite—investigators were able to determine the distribution of helium 3 in the body of a meteorite. "Contour maps" prepared by J. H. Hoffman and Alfred O. C. Nier at the University of Minnesota show that the helium-3 content falls off toward the center of the bodies, as would be expected. The contours also show that the

Grant meteorite reached the earth in something like its original shape, though it lost about half of its original mass to erosion by friction in the earth's atmosphere. In the case of Carbo the helium-3 contours indicate that most of one half was eroded away. Edward L. Fireman of Harvard University has measured the helium-3 content in Carbo with essentially the same results.

Helium 3 thus yields clues to the age and biography of the meteorite as a fragment of some larger parent body. But it does not reflect the age of the stone or iron that constitutes the meteorite. This is a more general and significant question because it bears upon the age of the solar system as a whole. If all the cold bodies in the solar system cooled down and crystallized at about the same time, as other considerations suggest, then a meteorite is as good a sample for this determination as a terrestrial rock which dates its crystallization from the most primitive era of our planet. Many meteorites have accordingly been subjected to the standard uranium-clock measurements. But stone meteorites also contain argon, which provides an independent check on the uranium clock.

Argon is the daughter of potassium 40, the relatively rare radioactive isotope of this common mineral element. The half-life of potassium 40 is 1.3 billion years, that is, half of the atoms of the isotope present in the rock disintegrate spontaneously in that period. The reaction yields two daughter elements: 89 per cent of the potassium 40 turns into calcium 40; the other 11 per cent turns into argon 40. As a consequence any mineral containing potassium gradually accumulates argon 40 in its lattice. So long as the system is undisturbed, the rate of argon accumulation is definite and immutable, governed only by the potassium content and by the rate of decay. By measuring both the potassium-40 and the argon-40 content in such an undisturbed sample it is possible to compute how long the system has been accumulating argon or, in other words, to calculate the date at which the sample crystallized.

This method was first applied to meteorites in 1951 by E. K. Gerling and his coworkers in the U.S.S.R. Since then investigators at various laboratories have dated many stone meteorites. A consistent pattern has emerged from the results: the potassium-argon ages of the most common stone meteorites tend to clump in the range of 4 to 4.5 billion years. Some of the ages fall below this clump, but none are higher. These find-

ings are consistent with the age of 4.6 billion years indicated by the lead-to-uranium ratios and the parallel decay of rubidium to strontium. That the potassium-argon age is somewhat "younger" is taken to reflect the fact that the parent body of the meteorites was hot early in its history; some of the argon would thus have been lost before the body cooled to a temperature at which it began to retain argon. The still younger ages found for some meteorites indicate either that the potassium occurred in minerals unfavorable for argon retention or that some argon may have boiled out of the meteorite on an unusually close passage around the sun.

An age of 4.6 billion years is now generally accepted for both meteorites and the crust of the earth. All evidence indicates that the stable, cold objects in the solar system took their present form at about that time.

Before the advent of the noble-gas technique for measuring time, it was quite difficult to go back further into the past with anything like certainty. The success of the technique in recent years encouraged investigators to apply it to the task of dating the origin of the elements themselves. With that ultimate starting point established, it might be possible to learn more about events in the period during which the solar system took shape from primordial matter. Inquiry into the age of the elements, however, takes one almost immediately into the thick of the controversy that divides cosmologists into two camps: the proponents of the "evolutionary" cosmology, who hold that the heavier elements were formed (along with hydrogen) all at once; and the "steady state" theorists, who contend that the formation of elements (including hydrogen) has been going on all along.

The simpler choice is to assume that the heavier elements were all created at once, or in a very brief period. One can then ask: When were they formed? Or, speaking of meteorites specifically, what was the time interval between the formation of the elements they contain and the time at which they crystallized?

One of the most important keys to the answer is the abundance ratio of the two long-lived radioactive isotopes of uranium—uranium 235 and uranium 238. Of the two, uranium 235 is much shorter lived, having a half-life of 700 million years, compared to 4.5 billion years for uranium 238. The relative abundance of these two isotopes is about 1 to 137. Going back in time, it is ap-

parent that the ratio has doubled about every 850 million years. If the original ratio was unity, as seems reasonable to assume, then these isotopes were formed no more than approximately 6.6 billion years ago. The very existence now of uranium 235, with its shorter half-life, places an upper limit on the age of the elements in the solar system.

The difficulty is that the original abundance ratio is not known. For each uncertainty of two, that is, with each doubling of the ratio, there is a corresponding uncertainty of about 850 million years in the time interval between element formation and the crystallization of the planetary bodies. The problem may be illustrated by a simple analogy. Suppose a clock is provided with a counter to indicate the number of times it has struck 12. The reading on the counter then tells how long the clock has run. But the reading has an inherent 12-hour error, because one does not know the hour at which the clock was set when it was wound. Under such circumstances another clock, wound at the same time but running, say, 40 times faster, would be highly useful. If the counter of the first clock were to read 1, the counter of the second would show something like 40. The original setting of the first clock could then be calculated with an error of only 18 minutes (one 40th of 12 hours).

Harrison Brown, now at the California Institute of Technology, suggested in 1947 that one of the shorter-lived radioactive elements might be used in just this way—as a faster-running clock to calibrate the original setting of the uranium clock. The original element would now be extinct, having disappeared from the solar system long ago by virtue of its relatively rapid decay; the clock would have by now completely run down. But it might still have been running at the time the meteorites were formed. In this case some atoms of the element would have been incorporated in the meteorites, and, under favorable conditions, they would have left a detectable "fossil" there. This fossil would be the daughter isotope to which all the radioactive material would have eventually decayed.

Following Brown's suggestion, the table of elements was searched for possible parent-daughter pairs. Since the parent isotope was necessarily extinct, it had to be found among the "artificially" radioactive products of the nuclear reactor or the high-energy accelerator. The most likely pair proved to be iodine 129 and its daughter xenon 129. A fairly

common fission-product, iodine 129 was almost certainly formed in nature along with ordinary stable iodine 127, and in approximately equal abundance. Because it has a half-life of only 17 million years, it has completely disappeared. This half-life seemed to be in just the right range—long enough so that some of the isotope might have been incorporated in meteorites, but short enough to provide a clock that runs down about 40 times faster than the uranium clock. The daughter, xenon 129, also seemed ideal for the purpose. It is a noble gas, and that portion of it which arose from the decay of iodine 129 would show up in meteorites as an excess over the average ratio of xenon 129 to xenon 132. But when the first attempts in 1955 and 1956 failed to reveal such excess xenon 129, the prospects for finding extinct radioactivity seemed poor indeed.

Then in November of last year a small piece of very crumbly stone that fell in Richardton, N.D., in 1918 yielded the looked-for excess in our laboratory at the University of California. The mass spectrometer showed clearly that several of the isotopes, when compared to the reference isotope xenon 132, were in anomalous abundance in the Richardton meteorite, but the most striking excess was that of xenon 129 [see bottom illustration on page 454]. This effect has since been confirmed in samples of Richardton studied at the University of Minnesota and at Heidelberg University, and in other meteorites studied at the University of California. In the black stone Indarch, which fell in 1891 in Trans-

caucasia, the xenon-129 peak is 3.4 times higher than the xenon-132 peak, a reading that is five times as striking as that in Richardton.

Quantitative studies of Richardton samples have already yielded preliminary figures. The amount of excess xenon 129 and the amount of stable iodine 127—which serves as an indication of the amount of iodine 129 originally formed—have been measured. The ratio of these two quantities is .000010. Calculating from the 17-million-year half-life, the time required for the iodine-129 to iodine-127 ratio to fall from 1, or thereabouts, to .000010 proves to be 290 million years.

This interval is not greatly affected when it is recalculated on the assumption (which accords with steady-state cosmology) that the heavier elements were built up over billions of years. William A. Fowler and Fred Hoyle at the California Institute of Technology and A. G. W. Cameron of the Chalk River Laboratory in Canada think it likely that a long succession of exploding stars have gradually built up the present inventory of these elements. Their estimates of this buildup period, based on theoretical production rates of the two uranium isotopes and of thorium, run close to 10 billion years. In this case the ratio of iodine 129 to iodine 127 at the conclusion of element formation in the vicinity of the solar system would not be 1 but .0025. Steady decay of iodine 129 over the long buildup period accounts for the large decline in the ratio. The time required thereafter to bring the ra-

tio down from .0025 to .000010—that is, the time from the formation of the elements to the incorporation of the elements in the solid bodies of the solar system—would be 137 million years. If iodine production was not uniform over the 10 billion years of element formation, but declined throughout this interval as some astrophysicists believe, another small correction must be applied—giving 120 million years instead of 137.

In terms of cosmological time the difference between 120 million years and 290 million years is not great. In either case it can be stated that there has been element-building which contributed to the solar system within the last 4.9 billion years.

The iodine-xenon clock is not yet fully calibrated. Past experience with other important radioactive clocks has shown that the clock becomes more reliable as experimental techniques improve. The iodine-xenon clock should ultimately provide a reliable time-scale for events which took place at about the time the solar system was formed. It is clear, however, that the time interval between element formation and the formation of the minerals in the meteorites is relatively short. Otherwise the iodine clock would have completely run down before the minerals were formed. The possibility that billions of years intervened between the formation of the elements of the solar system and the time its planetary bodies were formed is now conclusively ruled out.

The Author

JOHN H. REYNOLDS is assistant professor of physics at the University of California. He was born in Cambridge, Mass., in 1923 and attended Harvard University. After graduating in 1943 he served with the Navy for three years. From 1946 to 1950 he did graduate work at the University of Chicago, where he came in contact with Enrico Fermi, Harold C. Urey and Mark G. Inghram, the three men, he says, who most influenced his subsequent work. Since joining the faculty at California in 1950, his research interests have come to center on studies of isotopes in geologic and meteoritic samples. The results of his mass-spectroscopic studies of the Richardton stone meteorite are discussed in the present article. Reynolds is now midway through a two-year research pro-

fessorship at California's Miller Institute for Basic Research in Science, which "has let me really bear down on meteorite research."

Bibliography

- DETERMINATION OF THE AGE OF THE ELEMENTS. J. H. Reynolds in *Physical Review Letters*, Vol. 4, No. 1, pages 8-10; January 1, 1960.
- HOW OLD IS THE EARTH? Patrick M. Hurley. Anchor Books, Doubleday & Company, 1959.
- ISOTOPIC COMPOSITION OF PRIMORDIAL XENON. J. H. Reynolds in *Physical Review Letters*, Vol. 4, No. 7, pages 351-354; April 1, 1960.
- NUCLEAR COSMOCHRONOLOGY. William A. Fowler and F. Hoyle in *Annals of Physics*, Vol. 10, No. 2, pages 280-302; June, 1960.

A NEW SCALE OF STELLAR DISTANCES

by O. C. Wilson

The distance of a star can be determined by comparing its apparent with its intrinsic brightness. A new index to intrinsic brightness has been found in the calcium absorption spectra of many cool stars.

The measurement of distance is still a fundamental problem at every level of astronomy. Outside the solar system only about 170 celestial objects—the stars lying within some 30 light-years of the earth—have been located with an error of less than 10 per cent. Estimates for perhaps another 600 are half as good, but the great majority of them are stars of a particular type—intrinsically faint objects still in an early phase of their evolution. As to the remaining millions of observable members of our galaxy, only their average distribution is known with fair accuracy. We probably have a better idea of the distances to some galaxies a million or more light-years away than to most individual stars in the Milky Way.

Yet many questions about the structure and history of the galaxy can be answered only through the accurate placement of individual stars. Thus it is not surprising that astronomers are always on an eager lookout for improved methods of measuring stellar distances. Recently a new one has turned up. It applies to a substantial fraction of stars out to about 2,000 light-years, and it can fix their distance within about 15 per

cent. At the same time it has provided astrophysicists with the entertaining problem, still unsolved, of explaining why it works.

Every new astronomical yardstick must of course overlap an older one; eventually all are tied to the scale of distance for the closest stars. This is established by trigonometric parallax: the measurement of the apparent angular displacement of a nearby star, against the background of more distant ones, when it is observed from opposite sides of the earth's orbit around the sun. The method is direct and unambiguous, but it is limited by the small size of the angles that must be determined. At a distance of about 30 light-years the shift is only 1/36,000 of a degree—the apparent angular size of a penny seen at a distance of 24 miles! Even such an angle is measurable within an error of 10 per cent. At greater distances the attainable accuracy rapidly drops off, and at 60 light-years the apparent shift is essentially too small to measure.

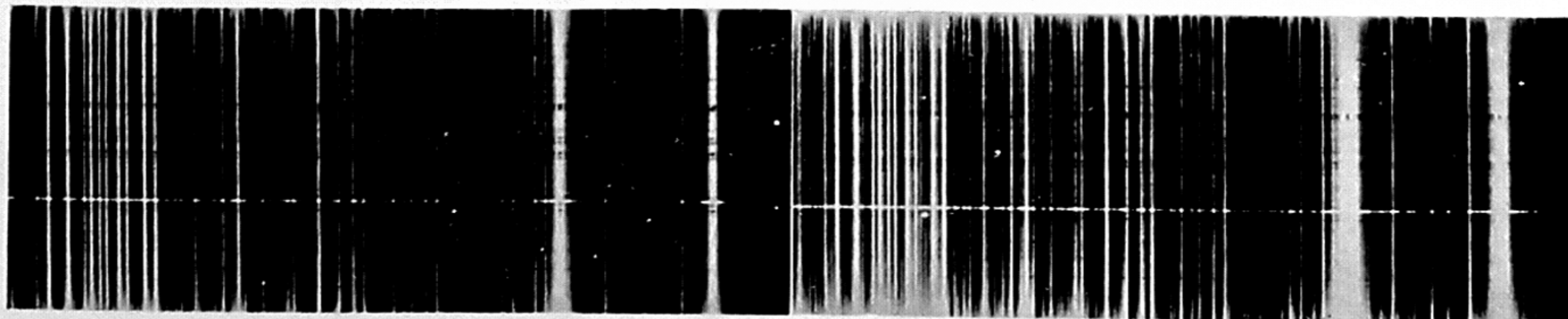
Beyond the effective range of trigonometric parallax it is no longer possible to get at the distance of individual stars directly. The only way to place them is

through a knowledge of their intrinsic brightness, or "absolute magnitude." The observed brightness of a star is proportional to its intrinsic brightness divided by the square of its distance. Thus determining the intrinsic brightness is equivalent to determining the distance.

Two ways of finding the absolute magnitude of certain types of stars have been known for half a century. One applies only to a class of pulsating stars: the cepheid variables [see "Pulsating Stars and Cosmic Distances," by Robert P. Kraft; SCIENTIFIC AMERICAN, July, 1959]. Although no one yet knows why, the rate at which the cepheids wax and wane is connected with their absolute magnitudes. The exact relationship is still in some doubt, but these stars have provided the best mileposts for the outer reaches of the Milky Way and for nearby galaxies. (The nearest cepheid is the Pole Star, 300 light-years distant.)

The second long-established index to intrinsic brightness bears the figurative name "spectroscopic parallax." The term refers to a correlation between brightness and the relative intensity of certain pairs of absorption lines in stellar spectra [see illustration on page 465]. Although the

K H



K AND H LINES of ionized calcium in stellar spectra form basis of a new method for measuring the intrinsic luminosity of stars. In this solar spectrum light and dark areas are reversed, and K and

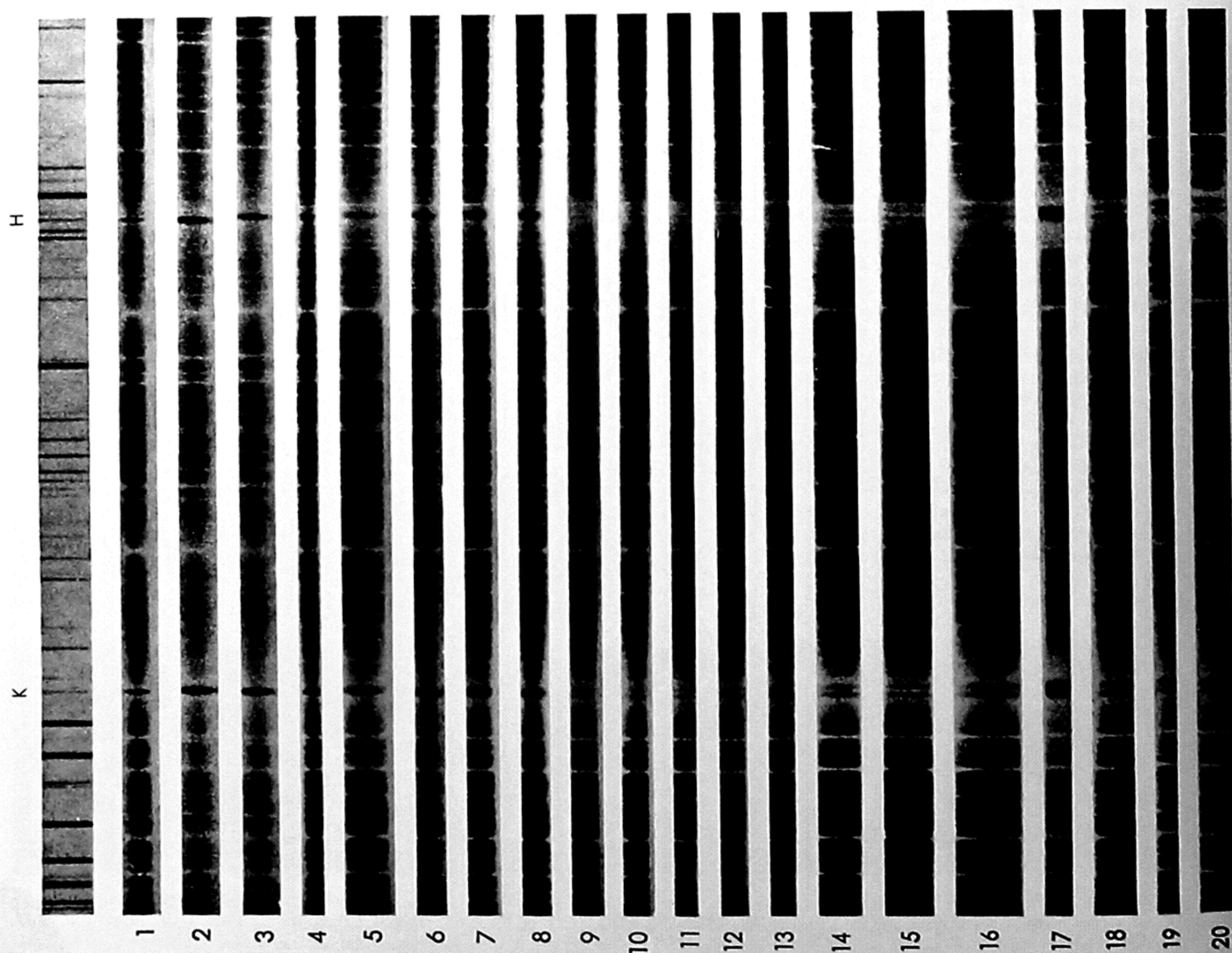
H absorption lines appear as bright bands. Black lines in their centers are "emission features," the widths of which are correlated with luminosity. Horizontal streak is caused by a sunspot.

details need not concern us here, this phenomenon does have an explanation. Briefly, spectroscopic parallax depends on the fact that larger, more luminous stars have relatively less mass, and therefore less surface gravity, than smaller ones. The resulting pressure differences in the atmospheres of the stars lead to differences in their spectra. When checked against stars whose distance (or absolute magnitude) is known from trigonometric measurements, spectroscopic parallax is found to be quite accurate for certain types of star. First devised at the Mount Wilson Observatory 50 years ago, and recently refined by W. W. Morgan and his associates at the Yerkes Observatory, the method has been very helpful in tracing out the major features of our galaxy within a few thousand light-years of the earth.

The new method is also spectroscopic, and it supplements rather than replaces spectroscopic parallax. It applies to stars of the spectral types designated G, K and M: the last three letters in the sequence—O, B, A, F, G, K, M—used to classify stars according to their surface temperature. The list runs from hotter to cooler: A typical O star might have a temperature of 30,000 degrees centigrade; an M star, 3,000 degrees. The stars of the last three groups represent a substantial fraction of the stellar population. Our sun is a member; its surface temperature of nearly 6,000 degrees classifies it as G2, meaning that its spectrum fits into the sequence 2/10 of the way from G0 to K0.

Spectra of these cooler stars always contain many dark lines, chiefly the absorption lines of neutral atoms and singly

ionized atoms (atoms lacking one electron) of the common metals. They are formed in the following way. When we look at a star, we see into its atmosphere down to a considerable depth. The light we observe is a sort of average made up of the contributions from the hotter and brighter lower layers and the cooler and fainter upper ones. At most wavelengths the depth to which we can see into a cool star is limited by the negative hydrogen ions (hydrogen atoms with an extra electron) that are always present in its outer gases. They produce a general haziness or opacity. At the wavelengths of the absorption lines, however, the emerging light encounters an additional obstacle in the scattering produced by the metal atoms and ions. Thus within the section of the spectrum covered by an absorption line we cannot see as far



down as we can at other wavelengths. The higher layers that we do see are cooler, on the average, than those producing the radiation in adjacent parts of the spectrum. Therefore the lines appear dark across their entire width.

This is the pattern of all the spectral lines of cool stars, with two exceptions. It is in the exceptions that we have found our new distance scale.

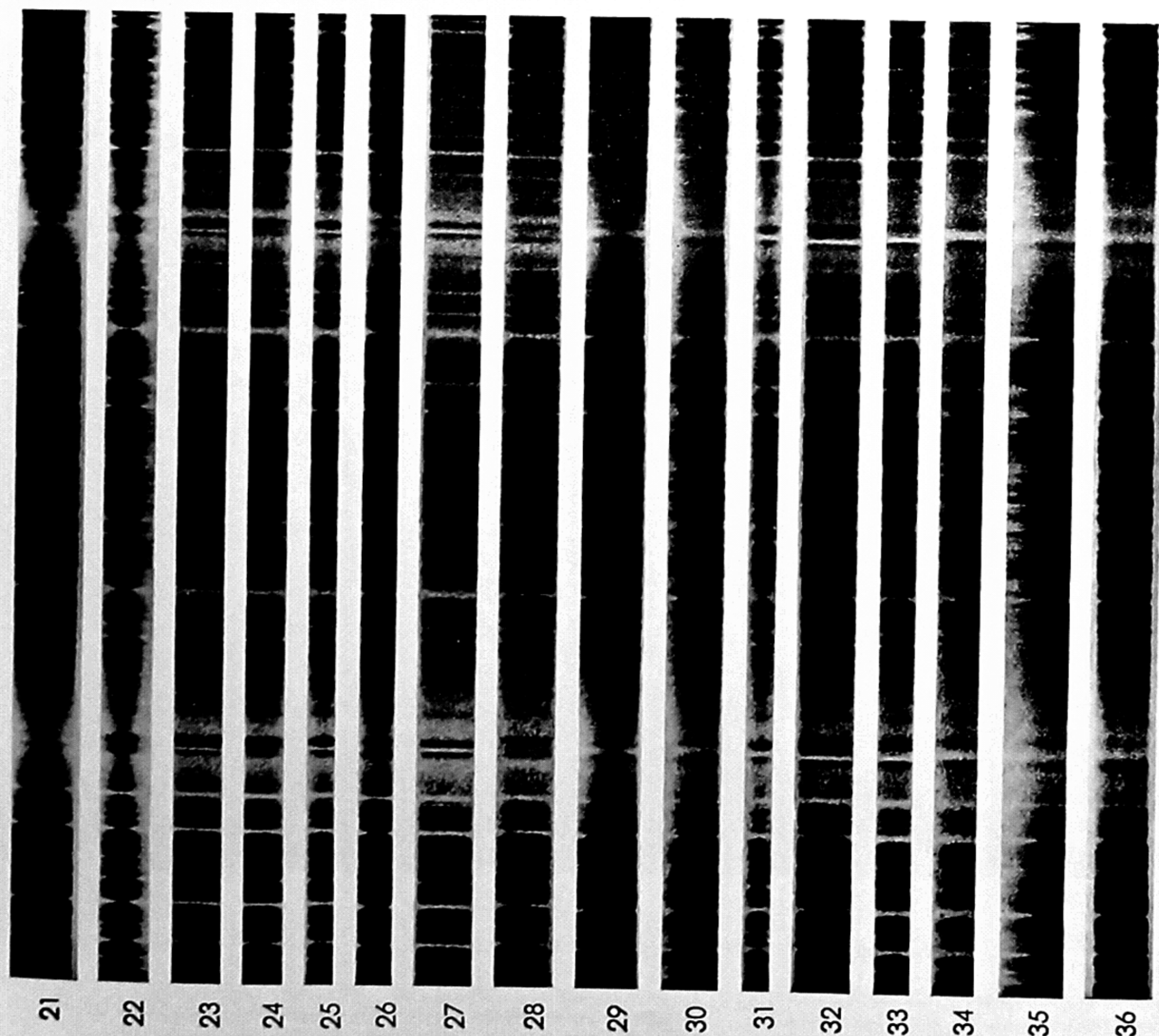
In the spectra of cool stars as seen at the surface of the earth the strongest absorption lines are a pair in the violet, designated H and K. (The system for naming spectral lines has no connection with the one mentioned earlier for distinguishing types of stars, although it uses some of the same symbols.) Produced by singly ionized calcium, the lines owe their strength to the fact that calcium is a fairly abundant metal which occurs

mostly in the singly ionized state in cool stars, and which has a high probability of making the atomic quantum-jumps associated with the H and K wavelengths.

What is remarkable about the lines, however, is not their strength but their internal structure. In the great majority of spectra of cool stars the H and K lines contain near their centers an emission component, or bright line, called a reversal. Almost always it is divided into two parts by a darker strip running down the middle [*see illustration at bottom of these two pages*]. The reversals in the solar spectrum, which are rather weak, have been known since about 1870, and were studied extensively at Mount Wilson in the early days of the Observatory. In fact the notation still

used to describe them was originated by George Ellery Hale, the first director of Mount Wilson. The broad absorption lines are denoted H_1 and K_1 ; the emissions, H_2 and K_2 ; the central dark band, H_3 and K_3 . In 1913 two German astronomers found the first H and K reversals in a stellar spectrum. Then observers began to discover the feature in one star after another, and eventually published extensive lists of the bodies in which they occur.

Even on spectrograms of quite modest dispersion, where the wavelengths are not spread widely, it is possible to see a systematic difference between the H and K reversals in intrinsically bright and intrinsically faint stars. In the latter each emission line is sharp and narrow, while in the bright stars it is noticeably wider, and the central dark strip is clearly ap-



LINES OF IONIZED CALCIUM are compared in the spectra of 36 cool stars. Intrinsic brightnesses increase from top to bottom, with a total spread of more than a million times. The central, light emission feature in the lines grows wider as luminosity increases. The correlation thus provides a measure of intrinsic brightness or distance.

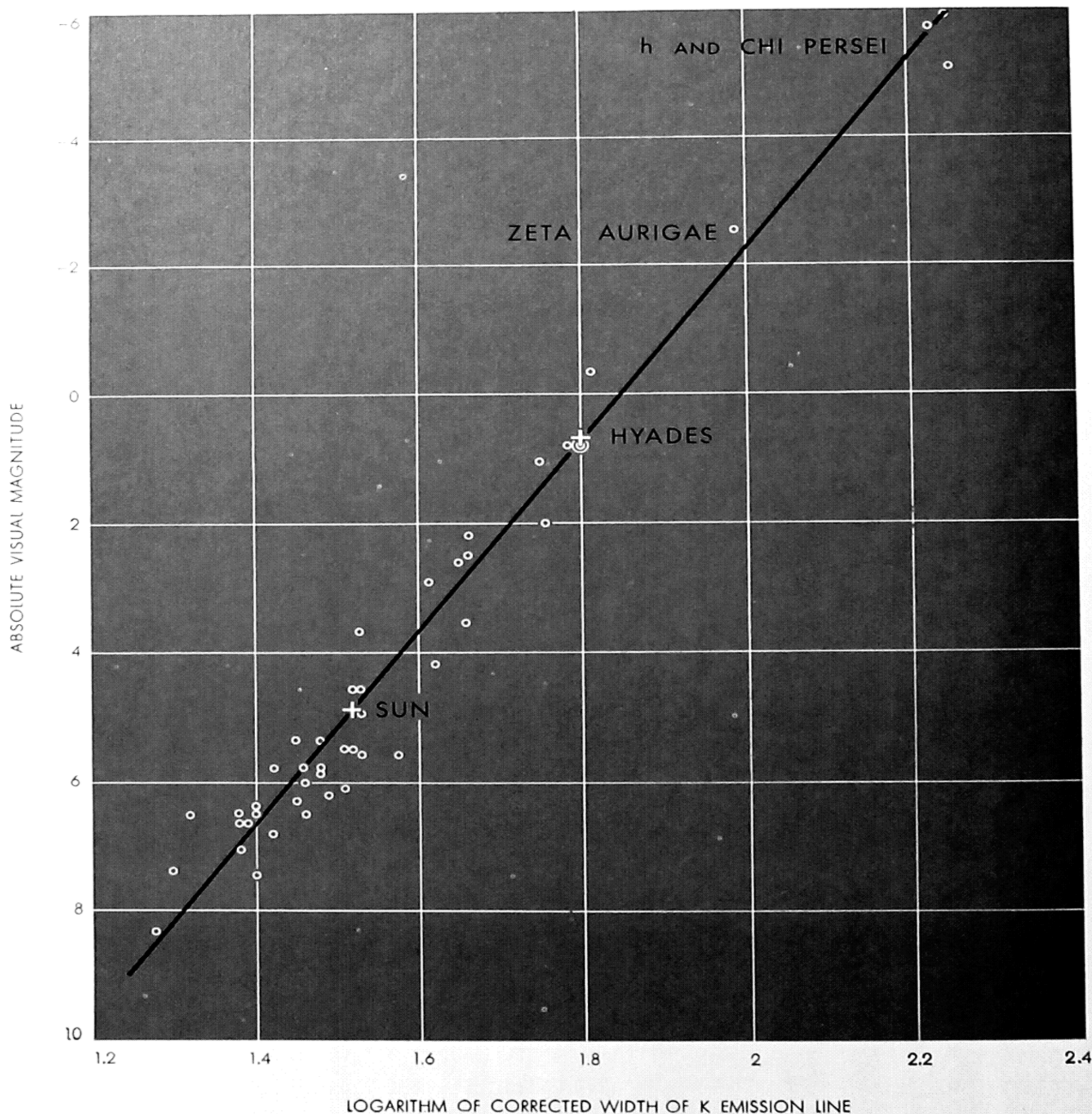
parent. Many astronomers must have noticed this fact, but curiously until a few years ago no one had followed it up.

At that time a young Indian astrophysicist, M. K. V. Bappu, and I decided to investigate the matter systematically. As test objects we chose a number of G,

K and M stars whose absolute magnitudes had been determined by the Yerkes group with spectroscopic parallax. With the 100-inch telescope on Mount Wilson and the 200-inch on Palomar Mountain we prepared spectrograms of these stars at a uniform dispersion. At

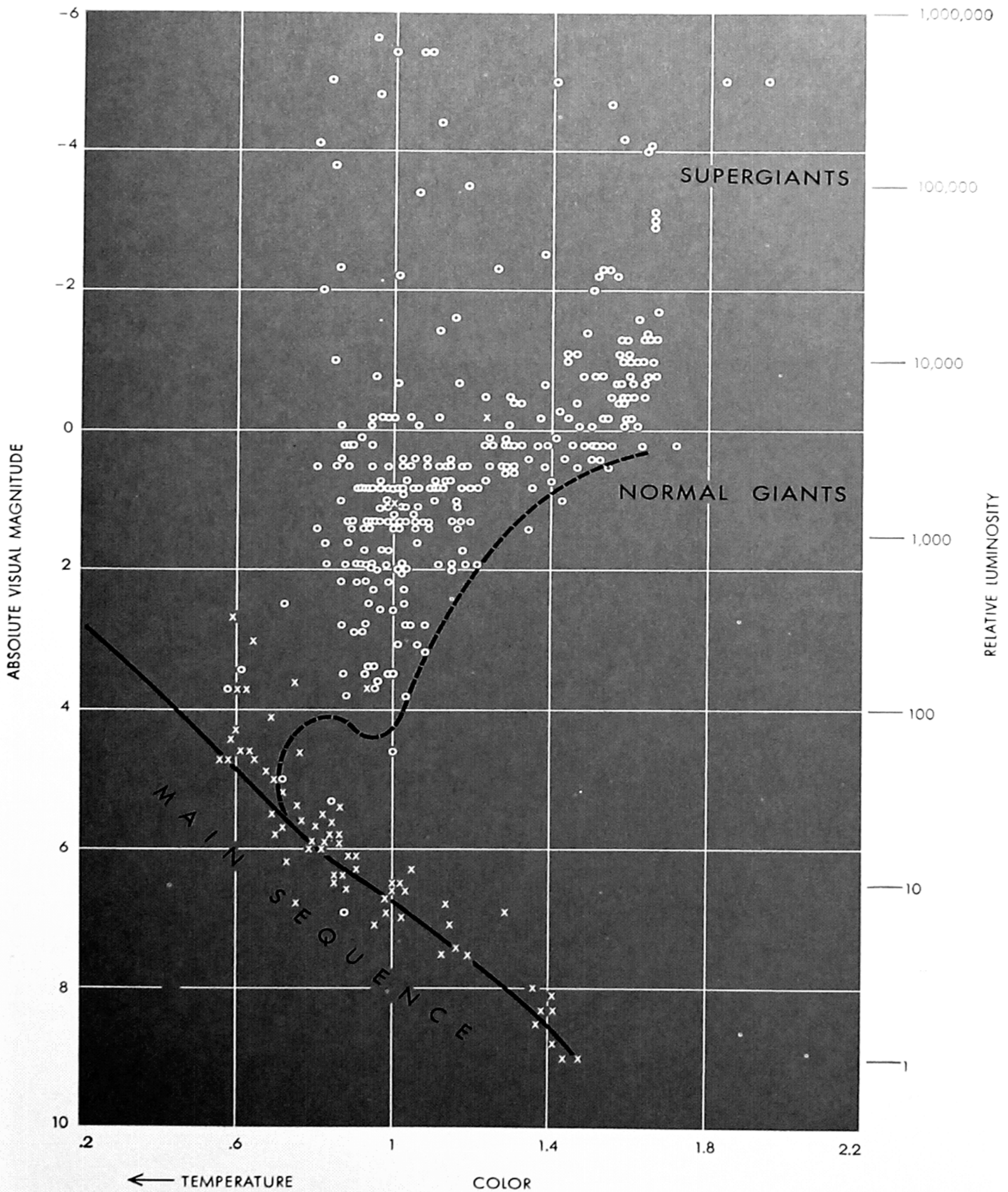
first it was not clear what aspect of the spectra we should concentrate on. After some preliminary trials the significant feature turned out to be simply the outside width of the bright components: H_2 and K_2 .

When the logarithms of the widths



CORRELATION between logarithms of width of K emission line and absolute magnitude is shown in this graph. (Negative values of the magnitude represent brighter stars than positive values.) The

straight line was drawn through points (crosses) representing the sun and stars in the Hyades cluster, whose absolute magnitudes are accurately known. Circles show other stars of known magnitude.



COLOR-MAGNITUDE DIAGRAM for stars in the neighborhood of the sun shows main-sequence stars clustering about solid line; giant and supergiant stars lying entirely above broken line. Mag-

nitudes found by K-line method appear as circles; those found by trigonometric parallax, as crosses. Temperatures increase from right to left, in the direction opposite to horizontal color scale.

were plotted against absolute magnitudes (which are also measured on a logarithmic scale), a dramatically simple relationship emerged. The points obviously clustered around a straight line running diagonally across the graph. The correlation extended over a range of 14 magnitudes (corresponding to a ratio of nearly a million between the luminosities of the faintest and brightest stars in this range). It apparently depended neither on the strength of the H and K emissions nor on the spectral type of the stars. Moreover, the nature of the scatter in the diagram suggested that much of it was due to a tendency of spectroscopic paral-

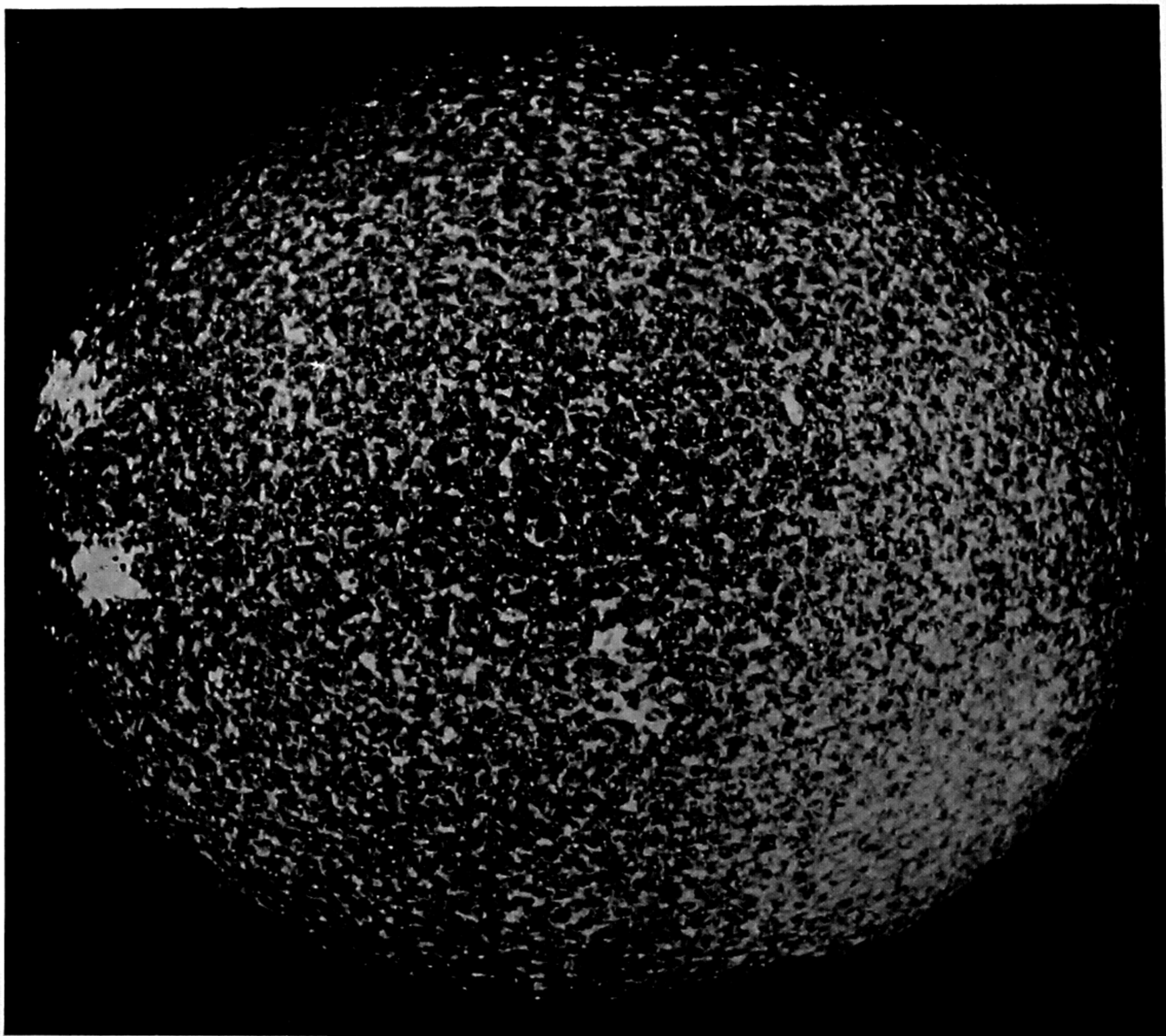
lax to lump together stars that actually differ appreciably in brightness.

Once the general shape of the relationship had been established, it was necessary to calibrate it more closely. Rather than fit a line to all the points, with their widely varying errors in the magnitude figures, we wished to place it by using only the most accurate available data. Since the relationship was linear, two points were in fact all we needed.

One of them was furnished by the sun. Its distance, and therefore its intrinsic luminosity, is known hundreds of times more accurately than that of any other

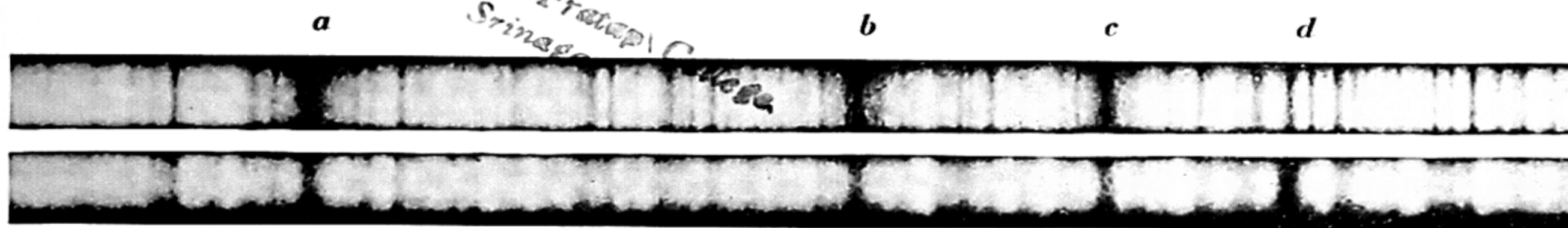
star. As has been mentioned, the H and K reversals in the solar spectrum are weak. On a high-dispersion spectrogram, however, their width can be precisely determined.

For the second calibration point we chose four K-type stars in the Hyades cluster. This is a physical grouping of stars, containing about 140 known members, which are close together and moving on parallel paths through space. (The brighter members of the cluster can be seen in the winter sky as a "V" in the constellation of Taurus.) Because it is the nearest such cluster in the heavens, 130 light-years away, the "proper mo-



SURFACE OF THE SUN is photographed in the light of the central emission of the calcium K line. Variations in brightness are prob-

ably associated with fluctuations in the strength of the magnetic field at the surface. The bright patches at left surround sunspots.



SPECTROSCOPIC PARALLAX measures intrinsic brightness of stars by comparison of the strengths of certain lines in their spectra. Upper spectrum is from an intrinsically faint star; lower one from a

star almost a million times as bright. Features marked *a*, *b* and *c* are absorption lines of neutral iron; *d* is an absorption line of ionized strontium. It is stronger, compared to iron lines, in brighter star.

tion" of its stars (their apparent motion at right angles to the line of sight) has been measured with exceptional accuracy. As with other fairly bright stars, their motion along the line of sight can be found from the Doppler shift in their spectra. Combining the proper and line-of-sight motions, and knowing that the stars are in fact moving along parallel lines, the distances to the individual members of the group can be calculated. The result is unusually precise for objects beyond the range of trigonometric parallax. When the distances are translated into intrinsic brightness, four of the K-type stars in the Hyades cluster are found to have nearly the same luminosity. Measuring the width of the H and K emissions in their spectra thus furnished the second calibration point. Now that the line was fixed, the absolute magnitude of any star could presumably be read off once the width of its H and K emission components had been measured.

To check the accuracy of the method, we applied it to a number of nearby stars whose absolute magnitudes were known from trigonometric parallax [see illustration on page 462]. A statistical comparison of the results showed that for so-called normal giant stars one good width measurement of the H or K reversal gives the intrinsic brightness with an uncertainty of about .3 magnitude (corresponding to an error of 15 per cent in distance). By using an average of several width measurements the uncertainty can be reduced to .2 magnitude. For dwarf stars the accuracy of the new method proved to be somewhat lower. This is not too serious because, as will appear in a moment, we are primarily interested in the distance of the giants. Among supergiants, the intrinsically very brilliant stars, good test objects are almost nonexistent. The binary star Zeta Aurigae provides one point, based essentially on spectroscopic evidence. Moreover, in the cluster known as h and Chi Persei, the distance of which can be determined on the theory of galactic rotation, there are several M-type super-

giants. Both Zeta Aurigae and the Persei cluster stars fit our line remarkably well.

Although there is no longer much doubt as to its accuracy, the H-and-K-line method has a serious drawback. It works only on stars with an apparent brightness sufficient to produce the necessary spectra in a reasonable time and with adequate dispersion. As a practical matter, this at present restricts observation to stars brighter than about the seventh magnitude. Soon the technique will undoubtedly be extended to fainter stars with the help of image-converter tubes or similar intensifying devices.

Even now, however, the technique can accomplish more than merely proving itself. In particular, it has already provided one of the best estimates of the age of our galaxy. Such estimates flow out of the modern theory of stellar evolution, which will be reviewed briefly here. The basic tool for analyzing the stages in the development of stars is the color-magnitude diagram: a graph in which individual stars are located horizontally according to their color and vertically according to their absolute magnitude [see illustration on page 463]. When the diagram is drawn up, it is seen that one class of bodies—smaller stars such as the sun—fall along a line known as the main sequence, running from top left to bottom right. To the right of the line in the upper part of the diagram are scattered the giants and supergiants.

The distribution is thought to come about as follows. Shortly after a star is born, it moves onto the main sequence, and remains there until a certain fraction of its hydrogen is consumed in nuclear reactions. Then it undergoes an internal rearrangement that converts it into a giant, moving it away from the main sequence upward and to the right. If a group of stars is formed at the same time, all its members will lie for a while on the main sequence. As the assemblage ages, the intrinsically bright stars, the spendthrifts at the top of the main sequence that burn their hydrogen most rapidly, are the first to move off. They

are followed by successively fainter stars as time goes on. At any given epoch the absolute magnitude of the faintest stars that have moved off the main sequence measures the age of the assemblage.

In the case of the stars of our galaxy we are doubtless dealing with objects of all ages, from the aboriginal inhabitants to some formed quite recently. Nevertheless, the faintest stars that have just left the main sequence furnish an index to the age of the oldest members of the collection. Hence by determining the absolute magnitudes of large numbers of the stars we can establish this lower border of the normal giants, which leads to the age of the galaxy. From the H and K measurements made thus far the age appears to be at least 10 billion years. As the boundary is sharpened by observations of more stars, and as the theory itself is refined, the estimate can be expected to improve.

A second important application of distance information is in finding the dynamic structure of the galaxy. As was mentioned in connection with the Hyades cluster, when the true motion of a star (or group of stars) is known, then the distance can be found from the line-of-sight velocity and apparent motion across the sky. Conversely, if we know the distance to a star, we can deduce its true motion from its line-of-sight velocity and apparent or proper motion across the sky. Thus the new technique for measuring distance will allow us to compute the galactic orbits of many more stars and to discover whether certain distant groups that appear to be moving together are actually doing so.

Finally we may ask how it is that a simple linear measure of a feature in the spectrum of a star can accurately tell us its brightness. Thus far there are only some glimmerings of the answer. Early studies of the H and K lines in the sun showed that the emission comes from a rather thin layer located above the region responsible for the ordinary absorption lines. Moreover, it was found that the atoms responsible for both the H₂ and K₂

emission are moving upward at about two kilometers per second, while those producing the central absorptions, H_3 and K_3 , are falling at a speed of about one kilometer per second. The reversals appear, at least weakly, virtually everywhere on the solar disk, but with many fluctuations of brightness and width over the surface. For instance, when the spectrograph slit crosses a sunspot, the emission lines over the dark central part of the spot are always much narrower than those over the normal surface.

In recent years observers on Mount Wilson have discovered a very strong correlation between the local magnetic-field strength and the intensity of the calcium emission over the solar surface. Evidently the H and K reversals depend somehow on the presence of a magnetic field. Perhaps hydromagnetic waves propel calcium ions, causing them to collide

violently enough with surrounding neutral atoms to excite the emission of the H and K lines.

The most likely source of energy for the hydromagnetic waves is the so-called hydrogen convection zone, a transition region lying not far below the visible surface in the cooler stars where the ionization of hydrogen changes from almost 100 per cent at the bottom to nearly zero at the top. As a consequence of the large difference in energy between ionized and un-ionized hydrogen, vertical convection currents are set up. They carry a large part of the outflowing stellar energy flux. Perhaps the velocities of the material in the convection zone are somehow transferred from the visible surface into the chromosphere, the layer immediately above the visible surface. Here their effects can be seen in the width of the H and K emission com-

ponents. Some theoretical work along these lines looks promising, but no really complete theory has been constructed, nor is there yet any general agreement among those who have thought about the problem.

Perhaps it is best to close by observing simply that, whatever the explanation of the correlation between emission-line width and luminosity, the fact that it holds over such a large range of stellar brightness implies a fundamental connection between the interior of a star, where its energy is generated, and the activity of the chromosphere. To find the nature of this connection is a problem of great general interest. In the meantime the correlation serves as a tool for attacking still other problems. It is one of those gifts that nature sometimes provides, as though in partial compensation for her obscurity elsewhere.

The Author

O. C. WILSON, a staff member of the Mount Wilson and Palomar Observatories, is an astronomer who accounts for his interest in the field by "an accident that occurred to someone I never met." After enrolling in a general science course in high school, he writes, "it developed that the regular teacher had broken his leg. He was replaced by a young and innocent substitute who spoke knowingly of various astronomical matters in such a manner that I was led to wonder if she really knew what she was talking about. In order to settle this point, I began reading books on the subject . . . and have been completely sold on it ever since. Unfortunately, the time interval is too great for me to remember now whether the substitute teacher was vindicated." Wilson acquired his B.A. at the University of California in 1929 and his Ph.D. in astrophysics at the California Institute of Technology in 1934. He joined the Mount Wilson Observatory as a computer in 1931 and has been there since that time (except for four years during World War II, when he did research on rockets and aircraft torpedoes at Cal Tech). At present all of

Wilson's research involves some form of astronomical spectroscopy; in particular the study of the H and K spectral lines discussed in his article.

Bibliography

- ACCURACY OF ABSOLUTE MAGNITUDE DETERMINED FROM WIDTHS OF H AND K EMISSION COMPONENTS. O. C. Wilson in *The Astrophysical Journal*, Vol. 130, No. 2, pages 499-506; September, 1959.
- A CENTURY'S PROGRESS IN DETERMINING STELLAR DISTANCES. Alfred H. Joy in *Astronomical Society of the Pacific Leaflet No. 173*; July, 1943.
- A COLOR-MAGNITUDE DIAGRAM FOR LATE-TYPE STARS NEAR THE SUN. O. C. Wilson in *The Astrophysical Journal*, Vol. 130, No. 2, pages 496-498; September, 1959.
- THE DISTANCES OF CELESTIAL OBJECTS. Suzanne E. A. van Dijke in *Astronomical Society of the Pacific Leaflet No. 204*; February, 1946.
- SOME THEORETICAL ASPECTS OF H AND K EMISSIONS IN LATE-TYPE STARS. F. Hoyle and O. C. Wilson in *The Astrophysical Journal*, Vol. 128, No. 3, pages 604-615; November, 1958.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

THE LIMITS OF MEASUREMENT

by R. Furth

The refinement of measuring devices has revealed a basic uncertainty of observation. An account of the physical basis of this limitation in the statistical nature of events at the molecular and atomic level.

SEVENTY years ago, when the system of ideas that we now call "classical physics" had been established, most physicists believed that they had before them an essentially true and very beautiful picture of the physical world. To perfect the picture, they thought, it remained only to put in more detail and make some minor corrections of the physical laws. For this it was necessary to increase the accuracy of quantitative observations, on the one hand by the construction of more sensitive and more accurate measuring instruments, and on the other by a very careful analysis of possible errors. Thus the experimenters of that period became primarily craftsmen in measuring techniques, and a physicist who succeeded in measuring a physical "constant" with a new degree of precision could be sure of making his way in academic life.

Nobody then doubted that the process of increasing the accuracy of observation could be continued indefinitely, at least in principle. Wonderful progress had been made in the construction of precision instruments and there seemed no reason why this steady improvement should ever come to an end.

The classical physicists realized, of course, that the process of measuring necessarily involves a mutual interaction between the object to be measured and the measuring device, and that this must to a certain degree change the measured object. For instance, when one uses a wire gauge to measure the thickness of a wire, the instrument causes a very slight deformation of the wire; one can avoid this by measuring the thickness optically under a microscope, but now the wire must be illuminated with concentrated light, which heats it slightly and makes it expand. Similarly, to measure an electric current one has to send at least a fraction of it through the measuring galvanometer, and that alters the original

value of the current. Yet it seemed that by increasing the sensitivity of the measuring devices one could reduce this interaction below any desired level, and that in any case corrections for the interaction could be calculated as long as the interaction was of a regular character and subject to known physical laws.

Around the turn of the century this and other basic problems in measurement began to become a great deal more involved, partly because of quite unexpected developments in experimental technique and partly as a result of new insight into the very nature of the laws of physics.

It is obvious that a physical quantity can be measured with accuracy only if it is precisely defined. Consider the measurement of the properties of a gas. According to the kinetic theory, the laws governing the "macroscopic" behavior of a gas hold strictly only for systems consisting of a theoretically infinite number of molecules. In other words, macroscopic quantities such as density, pressure and temperature are defined as statistical averages over infinitely large numbers of molecular events. But any device for measuring such a quantity can only record the outcome of a finite number of such events; for example, a pressure gauge records the net result of a large but finite number of molecular impacts on its sensitive surface. Hence even the most perfect instrument could not give the true value of the quantity to be measured, and repeated measurements will show irregular fluctuations, due to the irregular thermal movement of the molecules, which become more and more apparent with increasing sensitivity of the instrument.

THE CLASSICAL physicists did not believe that within the foreseeable future the sensitivity of instruments could be brought anywhere near the

level where these tiny irregular fluctuations would need to be seriously considered in measurement. But around 1905 Albert Einstein and the Polish physicist M. Smoluchowski independently showed that such fluctuations should under certain circumstances be observable by methods already available. They calculated that the fluctuations in the number of impacts of molecules on the surface of a small particle suspended in a fluid should produce an irregular movement of the particle, large enough to be observable and measurable under an ordinary microscope. The particle, in other words, acts as an extremely sensitive pressure gauge, and its movements record the fluctuations of pressure in the fluid. This phenomenon had actually been observed by many scientists; it had first been described by the English botanist Robert Brown in 1827. But it was not until almost a century later that the phenomenon was generally recognized as a manifestation of the thermal movement of molecules and given the name "Brownian movement."

The discovery of this phenomenon made it clear that there was no point in trying to increase the accuracy of measurement of such statistical quantities beyond the inherent statistical error. Indeed, when we try to do so, we find that we encounter another difficulty; the internal movement of the particles in the measuring instrument itself produces an irregular fluctuation of the readings, which increases with increasing sensitivity and increasing temperature. Thus it is useless to try to improve very sensitive torsion balances or radiometers by further increasing their sensitivity, for this at the same time also increases the fluctuations.

The same limitation applies to electric devices. In the first place, since electric charge is always attached to particles of finite mass (ions and electrons) and an

electric current is always connected with a flow of such particles, the current, charge density, voltage and so on must display irregular fluctuations when measured with sufficiently sensitive instruments. Secondly, the transfer of electricity through electronic tubes must be subject to fluctuations due to the "grainy" structure (*i.e.*, electrons) of the electric current. This is usually called the "shot effect," from a mechanical analogy: if you fill a funnel with small shot and let it pour out on a wooden plate, the flow of pellets will be more or less irregular, producing an irregular drumming noise. In communication systems using electronic tubes to amplify weak signals such fluctuations become apparent as background "noise." Clearly nothing can be gained in the performance of a receiver by increasing its sensitivity

beyond the level at which the fluctuations, or noise, begin to exceed the strength of the signals themselves.

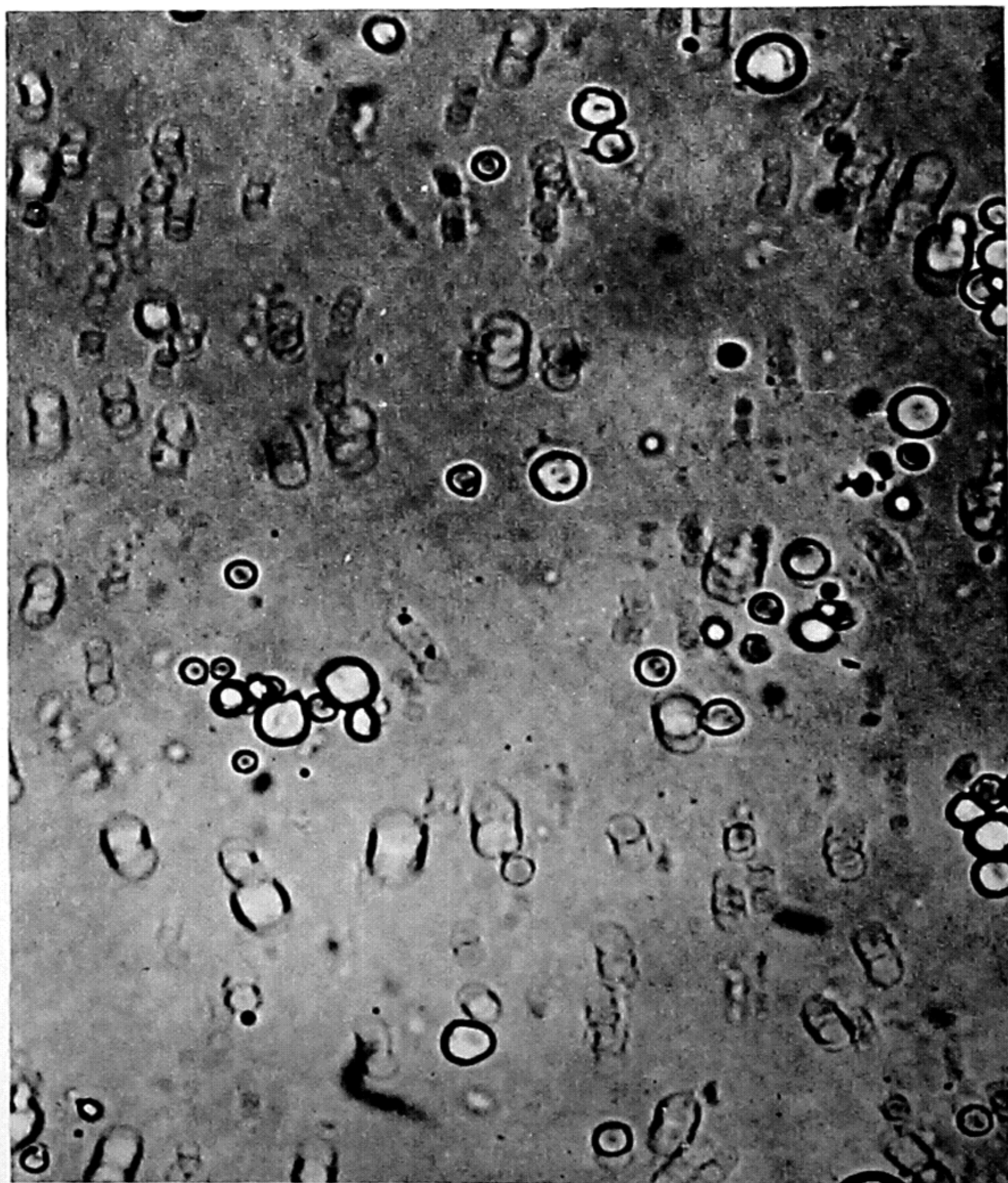
In view of these new facts the older ideas about the action of the measuring instrument on the measured object also had to be revised. Because of the Brownian motion of the measuring device this reaction has an irregular character, and therefore it cannot be completely corrected by compensatory calculations. For example, when a weak electric current is measured by means of a very sensitive coil galvanometer, inaccuracies arise from the inherent irregular fluctuations of the current itself, from the mechanical Brownian motion of the instrument, and from the motion of the coil within the magnetic field, which generates an irregularly varying current that is superposed on the current to be

measured.

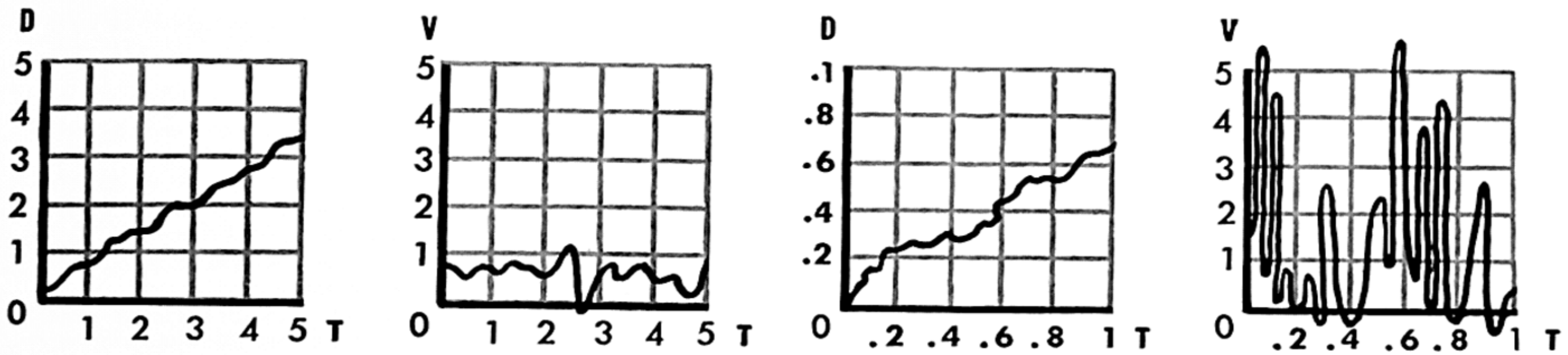
IN 1905 the Austrian physicist E. Schweidler called attention to another irregularity in nature. It had already been shown that the spontaneous decay of radioactive substances followed a very simple law: namely, that the number of atoms decaying per second is proportional to the number of undecayed atoms present. Schweidler interpreted this to mean that radioactive decay is a chance process; each atom has a definite probability for decaying within the next second, and this probability is independent of the atom's age, just as the probability for throwing a double six with a pair of dice is always one in 36 regardless of how long the game has been going on. Schweidler reasoned that if the decay of atoms does actually occur at random, the intensity of radioactivity from a piece of radioactive material should fluctuate irregularly in time. This was indeed proved experimentally soon afterward. It is now well known that the sequence of nuclear events is irregular in time. This holds not only for natural radioactive processes but also for nuclear processes produced artificially in the laboratory and for the mysterious processes responsible for cosmic radiation.

Thus it gradually became clear that nuclear processes were of a statistical nature and therefore incapable of being measured with infinite precision. In the meantime it had been found that the processes of emission and absorption of light could not be explained on the basis of classical physical laws, and quantum theory was introduced to overcome this difficulty. According to this theory, the emission and absorption of light by matter takes place in the form of finite quanta of energy or photons. It was Einstein who in 1917 suggested that the processes of emission and absorption were governed by probability laws like those governing radioactive decay, which implies that optical phenomena also are statistical in nature and cannot be measured with unlimited accuracy because of inherent fluctuations. This has been directly confirmed by experiment, for we now have devices of such extreme sensitivity that the statistical fluctuations of light intensity can be detected and even single light quanta can be counted.

All this makes one wonder if there are any physical quantities left which are not of a statistical nature. And what may be the origin of this chance character? To these questions Max Born in 1926 put forth a very radical answer which is now almost generally accepted. Briefly expressed, it states that the laws of quantum theory governing all atomic processes, and most probably all nuclear ones, are inherently statistical in charac-



THE BROWNIAN MOVEMENT of small particles is illustrated by this photomicrograph of oil globules in water. The multiple images are particles that have moved discontinuously under the thermal molecular bombardment.



INCREASED RESOLUTION of the movement of a particle under the microscope produces effects shown in these two pairs of diagrams. In first diagram the horizontal coordinate is time; the vertical coordinate, dis-

tance. In second the vertical coordinate is velocity. In third and fourth the magnification of the particle has been increased five times. Its position is now known with greater accuracy, but its velocity fluctuates more.

ter. They do not predict the fate of an individual atom or electron or photon but only give probabilities for the various events that may happen to it. This leads to the conclusion that all physical quantities, with the sole exception of the universal constants, are defined only statistically. Consequently they must exhibit fluctuations that will prevent precise measurements being taken beyond a certain limit.

SO FAR our argument on the limits of measurement has been more or less intuitive. Now let us analyze the problem a little more rigorously. It is very well known that the accuracy of measurement can be considerably improved by frequent repetitions of the measurement. Indeed, definite results may be obtained even from measurements that seem hopelessly erratic if only the number of observations is sufficiently large. The same result can be achieved by the use of recording instruments that measure the observed quantity continuously over a period of time, and by smoothing the graph obtained. Or the measuring device can be designed in such a way that it will automatically integrate the measurements over a certain time interval and so directly give the required time average. An example of one such procedure is the measurement of the mean intensity of a weak light source by long photographic exposure.

There is a very serious objection, however, to this kind of scheme: it presupposes that the observer already knows that the quantity to be measured is constant in time. But how is one to know this except by taking repeated measurements and comparing them to find out whether the value is maintained or not? And since the readings or records will vary in time because of the fluctuations, one will not be able to decide whether the quantity has really remained strictly constant or has varied within the limits of the recorded fluctuations.

Actually the principal task of the ex-

perimenter is to measure a certain physical quantity as a function of time. If he restricts himself to a short time interval, this becomes equivalent to the problem of determining, at one and the same time, a quantity and its rate of change as precisely as possible. The question to be answered is then: Is there a natural limit to this precision, imposed by the thermal fluctuations and by the statistical nature of the physical laws?

Let us first consider the effect of the fluctuations of the measuring instrument in this new light. Any measuring device can be characterized by its "relaxation time," that is, the time which must elapse between two successive observations to make them independent of each other. In the case of the human eye or ear, this relaxation time is about a twentieth of a second. In an ordinary mercury thermometer, it is of the order of several minutes; at the other extreme, in some electronic relays and counters, it is of the order of millionths of a second. Clearly when we use an instrument with a very long relaxation time for measuring a quantity, the accuracy of the measurement will be very great, because the instrument can integrate to even out the fluctuations, but no information will be obtainable about the rate of change of the quantity during this time interval. On the other hand, by using an instrument with a very short relaxation time we can repeat measurements at short intervals and so approximately measure the rate of change, but now the fluctuations will make themselves very strongly felt.

In 1933 the author showed that if one multiplies the mean error in the measurement of a quantity by the mean error of the simultaneous measurement of its rate of change in time, one obtains a constant, which is independent of the relaxation time. Thus an increase in the accuracy of the value for the quantity automatically diminishes the accuracy of the rate of change, and *vice versa*. Now this error constant depends on the construction of

the instrument and is proportional to the absolute temperature. So the only way to reduce the error arising from the thermal fluctuations of a given type of instrument is to lower its temperature.

WE NOW TURN to the question of the statistical fluctuations of the observed quantity itself. It will suffice to analyze in this light the most fundamental type of measurement; namely, the simultaneous determination of the position and the velocity, or rate of change of position in time, of a small particle in a fluid. We shall make the measurements with the help of a measuring microscope. The particle moves downward in the fluid as a result of the action of gravity. If there were no Brownian motion, the particle would move against the frictional resistance of the fluid with a constant speed, and the record of its positions over the time interval would be represented on a coordinate graph as a straight line. But because of the superposed Brownian motion the actual position-time record is irregular (*see first diagram at the top of this page*). In order to determine the average velocity of the particle from this diagram one can construct the velocity diagram (*second diagram*). This exhibits the spread of observational accuracy in the determination of the mean velocity. If one tries to increase the precision of measurement of the particle's position by using a microscope of greater magnification, one will obtain a curve (*third diagram*) which represents the beginning of the curve in chart A at higher resolving power. But now when one constructs the corresponding velocity diagram (*fourth diagram*), one realizes at once that what is gained in the accuracy of measurement of the particle's position is lost by the apparent increase in the spread of observed velocities from this average. This reflects the well-known phenomenon that the path of a Brownian particle seems to become more and more complicated the more one increases the

magnification of the observing microscope.

In practice the thermal fluctuations of both the observed and the observational systems and their mutual interaction have to be taken into account. It can be shown that the uncertainty of the measurement is again given by a relation of the same type as before; *i.e.*, the product of the errors in the measurement of a quantity and of its rate of change in time is again a constant. The over-all error is proportional to the average temperature of the three elements—the object, the measuring device and the coupling mechanism. In the simplest case, when the temperature is uniform throughout the three parts, they merge into one combined system, and it becomes impossible to attribute the observed fluctuations to any particular cause. For example, the fluctuations exhibited by a supersensitive galvanometer in measuring a current can be equally well attributed to its mechanical Brownian movement, to the electrical fluctuations in the connected circuit or to the fluctuating interaction between both.

To increase the accuracy as much as possible one will try to keep the average temperature as low as possible. But here again we soon reach a limit beyond which we cannot proceed without impairing the sensitivity of the method. Take again as an example the observation of the path of a small particle. Here the light scattered by the particle into the observing device (the eye or a camera) plays the role of the coupling mechanism. If we use strong light so we can see or photograph the particle under high magnification, we introduce a high-temperature radiation that increases the Brownian fluctuations. On the other hand, if we use a weaker light to reduce the temperature, we must go to lower magnification to see the particle and there is a loss of sensitivity.

SO FAR the analysis has been conducted along "classical" lines. We now have to revise it in the light of the theory of quanta. According to the classical conception, at the absolute zero point of temperature the thermal fluctuations should vanish. Quantum theory, however, implies that every physical system must retain a finite "zero-point energy," which means that however low the temperature, the system will continue to

fluctuate about an equilibrium position. Now if the system is bound to the equilibrium by strong forces, the fluctuations of position will be small, but it follows from the theory that the fluctuations in velocity will be large. On the other hand, if the system is bound loosely, the velocity fluctuations will be reduced but the fluctuations of position will inevitably increase. Quantitatively the product of the uncertainty of position and the uncertainty of velocity is approximately equal to Planck's universal constant of "action" (h) divided by the inertial resistance of the system. This relation is strikingly analogous to the corresponding "classical" relation. It shows that if the inertia of the measuring instrument is reduced sufficiently, then the inaccuracy of the simultaneous measurement of a quantity and its rate of change can be made very small, but it can never be reduced to nothing, even at zero temperature.

It is to be expected that these "zero-point fluctuations" will affect the behavior of the quantity to be measured as well as that of the measuring instrument. We can again restrict ourselves to the discussion of the movement of a particle under the action of some external force, say gravity. The classical view was that in the absence of thermal fluctuations the path of the particle should be completely determined, and its position at any time should be strictly predictable. But quantum theory holds that owing to the zero-point fluctuations this is actually not the case. In the words of Erwin Schrödinger, we may say that the particle performs a kind of "wobbling motion" about its classical path. As a result the progress of an individual particle cannot be predicted strictly; we can only calculate probabilities.

ALL THIS is summed up in Werner Heisenberg's famous "uncertainty principle." Applied to the question of measurement, this principle can be expressed as follows: The product of the uncertainty of the position and the uncertainty of the velocity of a material particle at one and the same time is approximately equal to Planck's constant, h , divided by the mass, m , of the particle. This indeterminacy is inherent in the very nature of the quantum laws of motion. Thus no instrument can be designed that would make it possible to

measure the position and speed of a particle simultaneously with a greater accuracy than h/m .

Plainly the quantum-mechanical uncertainty relation is of precisely the same type as the one imposed by thermal fluctuations, except that the former is independent of temperature and that no distinction can be made between zero-point fluctuations of the measured object and measuring device and the unpredictable interaction between them. Both processes affect all measurements, but the actual limit of accuracy in a particular case is determined by whether the Brownian movement or the zero-point fluctuations factor predominates: in measurements involving particles of large mass or high temperatures the limit of accuracy is fixed by the Brownian movement of the particle, while in those involving fundamental particles of very small mass, such as electrons, protons and neutrons, or very low temperatures, the quantum-mechanical limitation will be decisive.

Indeed, in the case of electrons the quantity h/m is of the order of magnitude unity, *i.e.*, about 1. On the other hand, the diameter of an atom is of the order of 10^{-8} centimeters, and the velocities of electrons within the atom are of the order of 10^8 centimeters per second, so their product is also of the order of magnitude unity. In other words, the error involved in the simultaneous measurement of the position and velocity of the electron is of the same order as the quantity to be measured. It becomes clear at once that observation of the movement of the electron within the atom is out of the question; the observational method is so crude that it completely upsets what it endeavors to investigate. The same is true of the investigation of intranuclear processes by means of the bombardment of nuclei, and of the observation of the effects of this bombardment.

To summarize we may say that the aim of quantitative observation is the simultaneous measurement of physical quantities and their rate of change, and that the accuracy of this process has a natural limit. This limit is imposed both by the chance character of the thermal agitation within the instrument and the measured object, and by the statistical nature of the quantum-mechanical interaction processes among the fundamental particles.

The Author

R. FURTH is reader in theoretical physics at Birkbeck College of the University of London.

Bibliography

NATURAL PHILOSOPHY OF CAUSE AND CHANCE. Max Born. Oxford University Press, 1949.

THE PHYSICAL PRINCIPLES OF THE QUANTUM THEORY. Werner Heisenberg. University of Chicago Press, 1930.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

THE MASS SPECTROMETER

by Alfred O. C. Nier

This sensitive instrument sorts atoms and molecules of various weights by means of a magnetic field. Developed by physicists, it is now widely utilized in other sciences and in technology.

THE MASS spectrometer is a classic example of the instrument that starts as a laboratory contrivance and becomes one of the great general tools of mankind. It grew out of the electrified glass tubes with which physicists liked to experiment in the late 1800s; today it is used as a research instrument in fields as widely diverse as chemistry, geology, biology and medicine, and is an indispensable piece of equipment in various industries. As recently as 1940 there were probably fewer than a dozen such instruments in operation in the

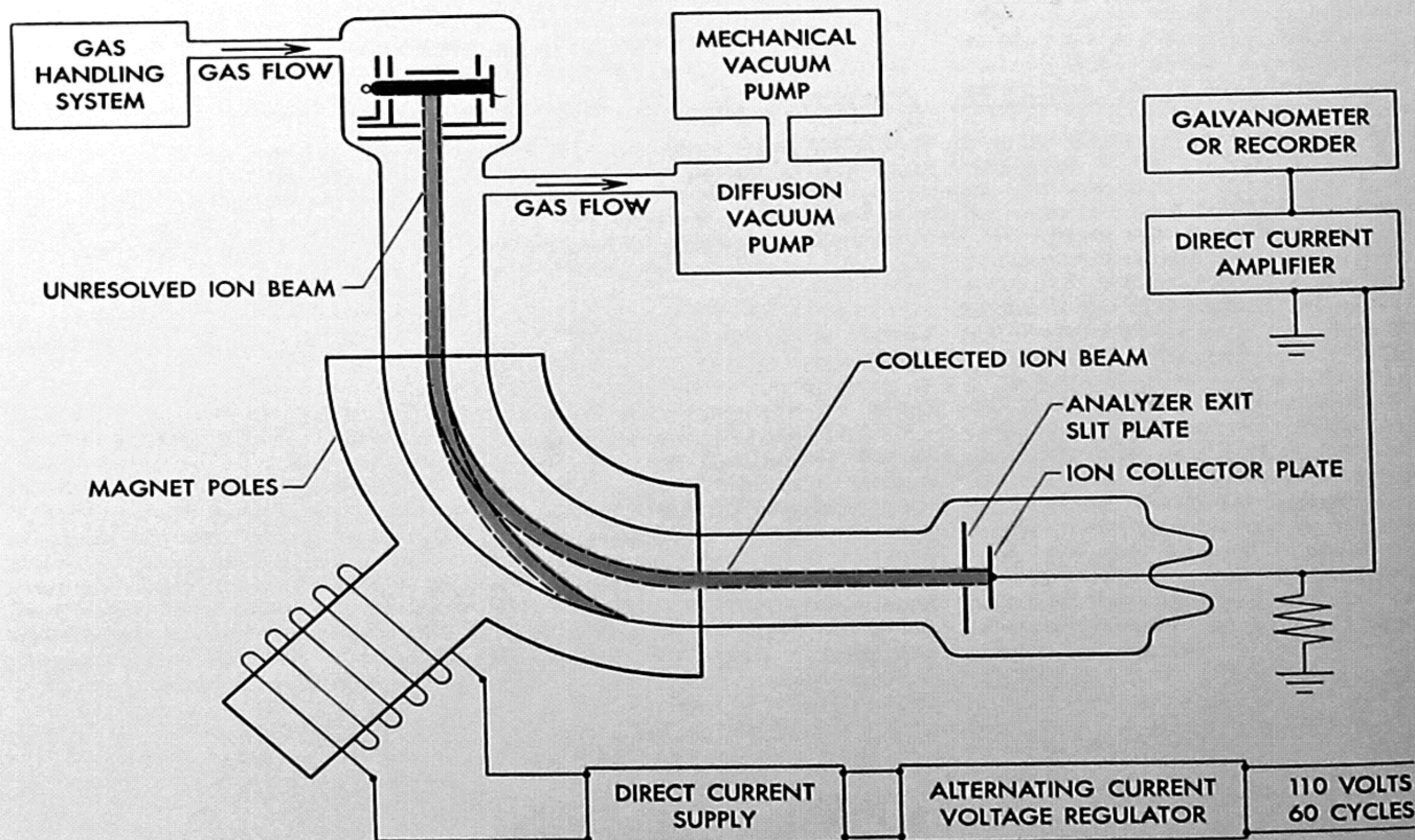
world; now there are many hundreds, and their use steadily grows.

Basically the mass spectrometer is an instrument for sorting and identifying atoms or molecules. The particles are ionized (given an electric charge) and then fired across a magnetic field which bends their flight into a circular path. The heavier of two ions will fly in a wider arc than the lighter one and hit a different target. The instrument readily separates ions with very slight differences in weight. And since the speed of the particles, strength of the magnetic

field and paths of flight are all accurately known, it is easy to calculate precisely the relative masses of the various ions.

A mass spectrometer sorts a mixed stream of ions by atomic weight just as a glass prism separates a beam of white light into a spectrum of its component colors or wavelengths of light. If the sorted ions are caught on a photographic plate, the apparatus is called a mass spectrograph; if they are detected and recorded by electrical means, it becomes a mass spectrometer.

The instrument can easily identify the



SCHEMATIC DIAGRAM of a mass spectrometer shows path of ions (in color). Those that reach the collector plate signal their presence by a minute elec-

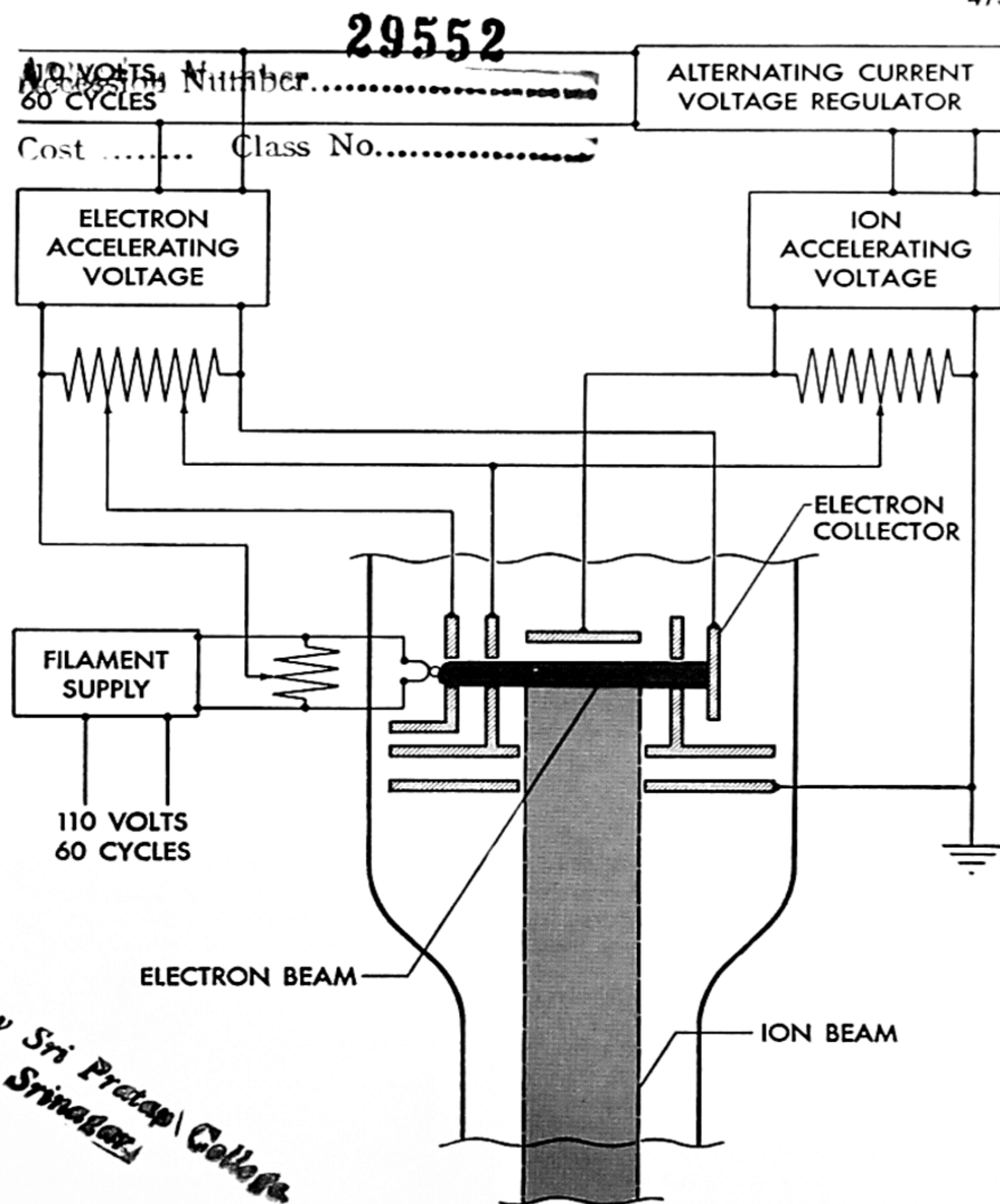
tric current. Adjusting the voltage that speeds ions down the tube, or the magnetic field that bends their path, brings ions of desired weight to the collector.

ingredients of air or any other gas and offers almost unlimited possibilities as a gas analyzer. It can analyze solids when they are converted to the gaseous form.

In principle the mass spectrometer goes back, as do so many powerful instruments today, to humble beginnings in the observations of research workers in the latter part of the last century. Physicists then were greatly interested in the study of electrical discharges in gases. To perform these studies they used a glass bulb having a pair of sealed-in metal electrodes, called anode and cathode. When a high voltage was applied across the electrodes and pressure in the bulb was reduced by pumping out some of its air or gas, the remaining gas became a conductor and glowed like a neon sign. From observations of this glow came some of our most fundamental and now familiar concepts of the nature of matter.

CHIEF among these were the discoveries of J. J. Thomson, who in 1897 began a famous series of experiments. Investigating the then mysterious cathode rays, which streamed from the negatively charged cathode in such a tube, Thomson determined that they were made up of particles. The individual particles always had the same mass and electric charge, regardless of the gas used in the discharge tube. Thus he discovered the electron, the smallest particle of matter, and opened the electronic age. He then turned to the anode rays seemingly coming from the positively charged anode. These also, he demonstrated, were streams of particles, but in their case the mass of the particles depended on the atomic weight of the gas in the tube. (Actually his anode rays were positively charged ions—atoms or molecules stripped of an electron by collisions in the electrical discharge.) Thomson thereby contributed to the growing chain of evidence that led to the realization that all matter is composed of atoms and that the weights of the atoms are directly proportional to what chemists had long recognized as the relative combining weights of the elements.

Thomson passed the positively charged anode rays through suitable combinations of electric and magnetic fields and found that the rays could be separated according to mass as the ions of different weights were pulled into different paths by the field forces. And in further exploration of the atomic weights of elements by this means he made another discovery. Chemists, in setting up their atomic weight scale, had arbitrarily taken oxygen as 16. The relative weights of the other elements tended to be integers also. But there were well-known discrepancies from this whole-number rule. For example, the rare atmospheric gas neon had an atomic weight of 20.2. In 1912 Thomson put a bit of pure neon



DETAIL OF ION SOURCE shows electron beam that ionizes the gas to be analyzed and the slotted electrodes through which the ion beam emerges. Voltages applied to the electrodes determine speeds of electrons and ions.

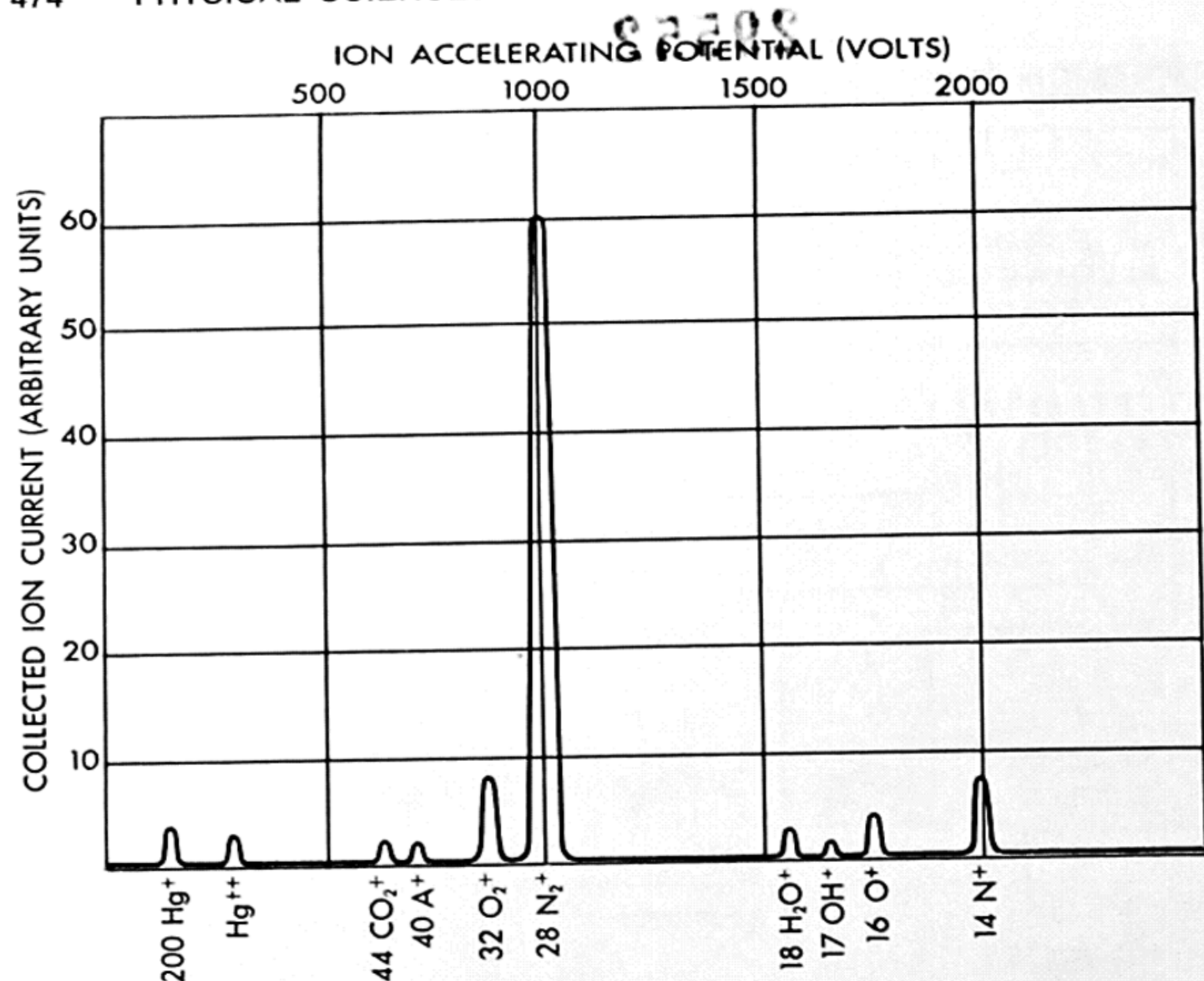
in his apparatus and found that it consisted of not one type of atom but two, some with a weight of 22, many more with a weight of 20, averaging out to 20.2. Thus he established the fact that some elements have isotopes—atoms of slightly different weight but the same chemical nature.

Mass spectrometry flowered almost directly from the subsequent study of isotopes. In 1918 the late Arthur Dempster of the University of Chicago built the first mass spectrometer. He designed the instrument to measure the relative abundance of isotopes in various elements. A year later the English physicist F. W. Aston, a colleague of Thomson, made a mass spectrograph and began a systematic 20-year study of isotopes in the entire atomic table.

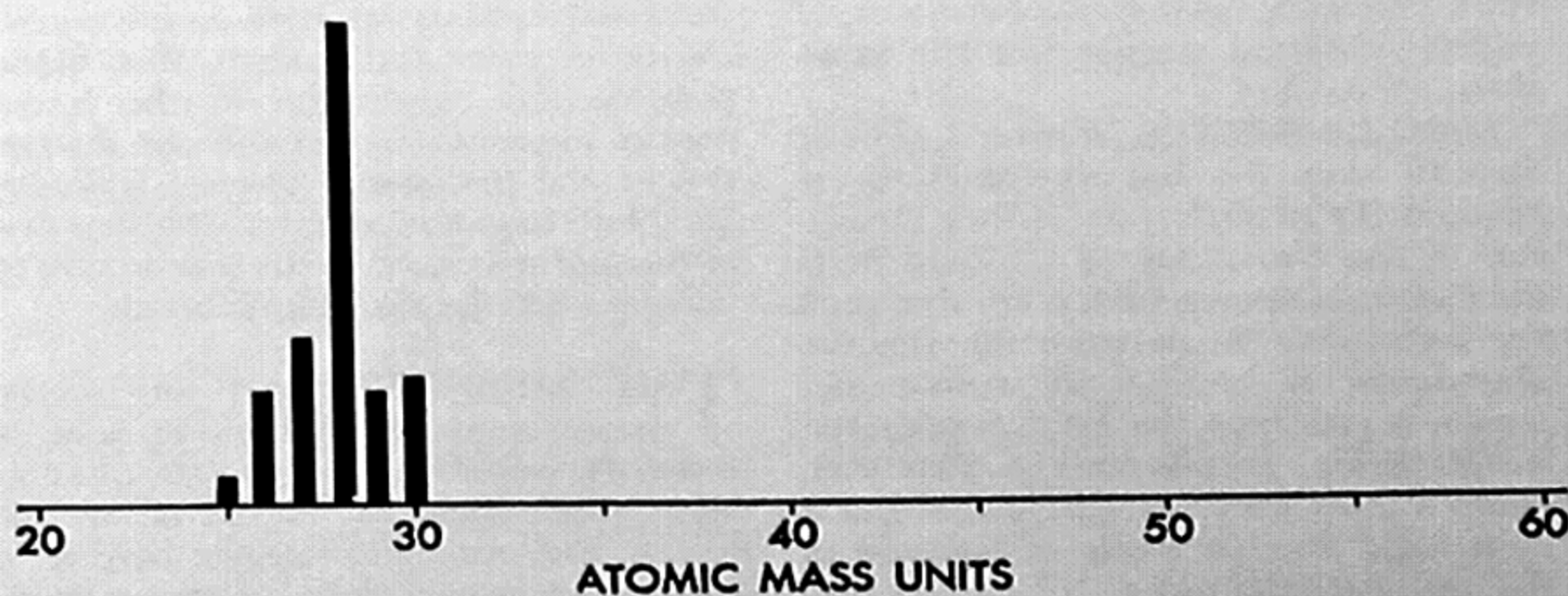
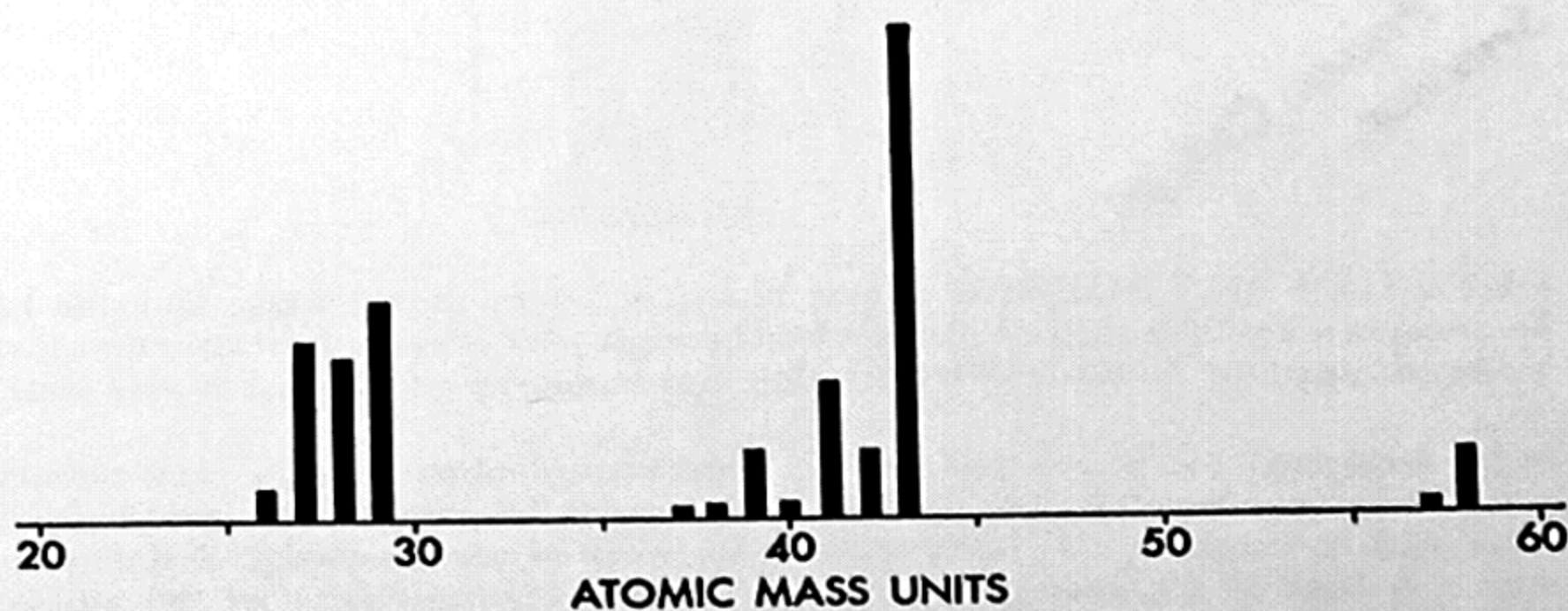
Although many of Aston's measurements of minute deviations in atomic masses are now superseded by more ac-

curate ones, obtained with more refined apparatus, his patient work laid the basis for much of our knowledge of the structure and binding forces of the atomic nucleus. It was this work, combined with Einstein's famous equation on the equivalence of mass and energy, that made possible the calculation of the latent energy in atomic nuclei and the discovery of the fissionable isotope uranium 235. And the mass spectrometer was one of the methods used in the production of uranium 235 for the atomic bomb.

THE MEASUREMENT of atomic masses remains an exciting field of research. From it we should continue to learn much about the nature of nuclear forces. But our chief interest here is in the wider applications of mass spectroscopy. By now it has enabled us to measure the relative abundance of most, if not all, of the isotopes occurring in



SPECTRUM OF AIR is plotted in terms of accelerating voltage required to focus each ion. The electron bombardment has not only ionized the various air molecules but has split some of them apart, as is shown by presence of monatomic oxygen and nitrogen. Mercury comes from pump.



HYDROCARBON SPECTRA for butane (*top*) and ethane (*bottom*) indicate relative abundance of molecular fragments after gas is broken up by electrons in spectrometer. Each pure hydrocarbon gives a unique picture.

nature and of those produced artificially in the laboratory. The study of isotopic abundances has many valuable applications in other fields of science besides physics. In addition, the mass spectroscope's convenience as an analyzing instrument makes it an important industrial tool.

In the early instruments the gas to be analyzed was ionized by bombardment with electrons from an electrical discharge. Most modern mass spectrometers produce ions by a more stable and controlled means. In the commonest type of instrument, the electrons stream from a hot tungsten wire and are accelerated through a potential difference of about 100 volts to give them the kinetic energy required to ionize the gas molecules. The gas is admitted continuously into the vacuum chamber of the instrument through a small "leak." The chamber is continuously pumped so that the net pressure is reduced to about one billionth of an atmosphere—a pressure so low that once an ion is formed there is little likelihood it will collide with another ion or gas molecule before it is caught on the detecting plate.

The ions are fired into the magnetic field at high speed. This is done in one form of the instrument by directing them through slits in a pair of parallel metal plates, across which an electrical potential difference is applied. Emerging from the slit in the second plate, the ions coast in straight lines until they pass between the poles of the magnet. The magnetic force, exerted at right angles to their direction of motion, causes the ions to curve in circular paths. The radius of curvature depends on the number of electrons lost in ionization (usually one), the difference in potential through which the ions fell, the strength of the magnetic field and the atomic weight of the ions being measured. By adjusting either the magnetic field or the difference in potential, ions of any particular weight can be made to hit a predetermined target. This is a slit behind which is mounted a plate. The ions impinging on this plate set up a tiny electric current, which is measured by special instruments.

ONE of the practical applications of the mass spectrometer is in the petroleum industry. Here the molecules to be analyzed are organic gases made up of many atoms. When struck by the electrons, such molecules are shattered into ionized fragments. For a given energy of bombarding electrons, these fragments may be made to give a mass spectrum characteristic of the gas bombarded, down to the actual arrangement of atoms in the molecule and the number and types of atoms present. Thus mass spectrometry can distinguish between two or more different chemical compounds having the same molecular

formula. This is extremely useful in the petroleum field, where many compounds are so nearly alike in properties that it is often difficult to identify them in a mixture by ordinary chemical means. Many oil refinery laboratories have set up spectrum patterns for hundreds of pure compounds and run routine analyses of mixtures having as many as several dozen components.

Such analyses are usually made in laboratories on samples brought to the instrument. But a mass spectrometer can also be put directly into the process stream in a continuous-flow chemical plant to provide a continuous analysis and automatic feedback control of the composition of the process stream. The first notable application of this kind was in the U-235 gaseous-diffusion plant at Oak Ridge, where the process stream is continuously monitored by a battery of strategically placed instruments for traces of air, refrigerants and other contaminants that may leak into the system. Each instrument is pretuned to the impurities being watched, just as a push-button radio is pretuned to certain stations. A clock and switch arrangement puts the instrument through its preset tests in sequence. Each analysis takes 24 seconds, and a single instrument can make 3,600 determinations a day, using small samples of gas which total altogether only two cubic centimeters. The amounts of the impurities are automatically recorded on a single chart, which allows an operator to see the entire chemical analysis of the stream at any time and note any changes that have taken place in it.

A portable mass spectrometer was developed to detect minute leaks in the system. The plant has miles of pipes, valves, vacuum pumps and vessels that must be kept tighter than those in any ordinary chemical plant. To have checked the completed plant for leaks by conventional means would have required about 1,000 test stations, each manned by a team working 8,000 hours to cover the million or so critical points in the system. With mass spectrometers acting as leak detectors the gigantic task was accomplished with only about 3 per cent of that effort. The plant was completed with great savings in time, materials and skilled manpower.

IN BIOLOGY the mass spectrometer has been useful for studying respiration, among other things. F. A. Hitchcock and his colleagues at Ohio State University have used a continuous-reading instrument to analyze the composition of respiratory gases under various conditions, including sudden changes in pressure. At the University of Minnesota Fletcher Miller and Allan Hemingway analyzed, with a special portable instrument, the rate of nitrogen elimination in patients suffering certain lung disorders,

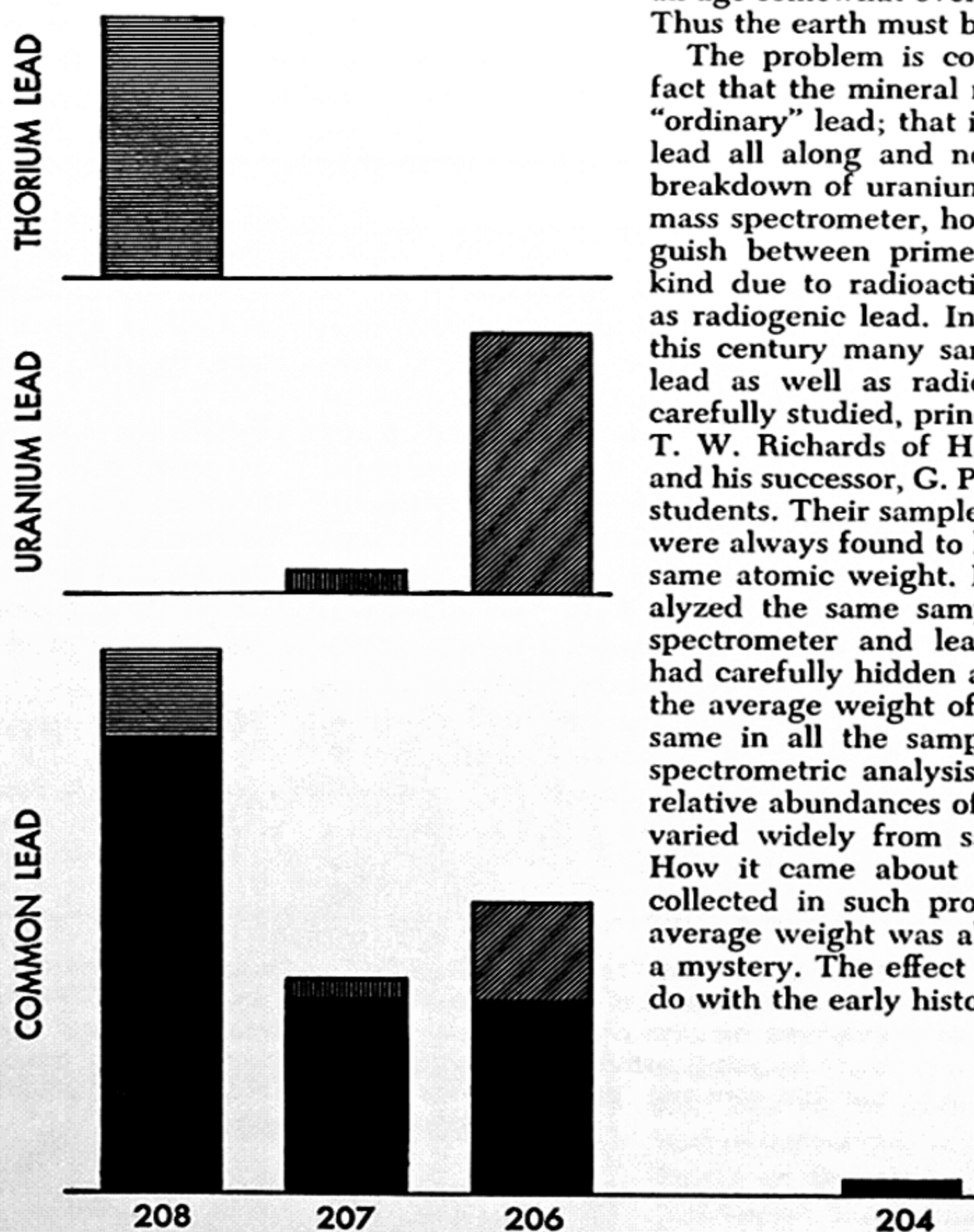
and other workers at the University have made continuous respiratory gas analyses during thoracic surgery. At the same institution the botanist Allan Brown and his students have examined metabolic gas exchanges in growing plants. Here the mass spectrometer follows continuously the composition of the gas surrounding the plants as the photosynthetic process takes place.

The mass spectrometer is an indispensable item of equipment in biological tracer experiments employing stable isotopes. For many investigations of metabolism stable isotopes are preferable to radioactive isotopes; for some they are the only ones available. For instance, nitrogen and oxygen, two important elements in biological studies, have no suitable radioactive isotopes but useful stable ones. The mass spectrometer can measure extremely small traces of an element; with it, for example, M. G. Inghram and his colleagues at the University of Chicago have determined within 10 per cent that the quartz in a sample of granite contains only one part of uranium in 10 million. Comparable

precision has been obtained in measuring other trace elements.

ONE OF THE most interesting uses of the mass spectrometer has been in reckoning the age of the earth. The method involves measuring the isotopic abundances of lead, the product of the breakdown of radioactive uranium and thorium, in the earth's rocks. In the long course of time uranium 238 decays to the lead isotope 206, and uranium 235 to lead 207. Thorium decays to lead 208. The rates at which these transformations take place are accurately known. Thus by making a chemical analysis of the amounts of uranium, thorium and lead in a sample of mineral, and determining the relative amounts of the three isotopes of lead, one can make three independent calculations of the age of the mineral. Once the ages of the uranium and thorium minerals are calculated, it becomes possible to estimate accurately the ages of the deposits in which they lay and thereby to set up an age scale for the various geological eras. The oldest minerals found so far have an age somewhat over two billion years. Thus the earth must be at least this old.

The problem is complicated by the fact that the mineral may contain some "ordinary" lead; that is, lead which was lead all along and not formed by the breakdown of uranium or thorium. The mass spectrometer, however, can distinguish between primeval lead and the kind due to radioactive decay, known as radiogenic lead. In the early part of this century many samples of ordinary lead as well as radiogenic lead were carefully studied, principally by the late T. W. Richards of Harvard University and his successor, G. P. Baxter, and their students. Their samples of ordinary lead were always found to have precisely the same atomic weight. But in 1937 I analyzed the same samples with a mass spectrometer and learned that nature had carefully hidden a secret. Although the average weight of the lead was the same in all the samples, the accurate spectrometric analysis showed that the relative abundances of the lead isotopes varied widely from sample to sample. How it came about that the isotopes collected in such proportions that the average weight was always the same is a mystery. The effect apparently has to do with the early history of the lead—its



SPECTRA OF LEAD from various sources show differences in isotopic content. Thorium decays to lead 208 (*top*); uranium 238 ends as lead 206; uranium 235 as lead 207 (*center*); natural lead (*bottom*) is a mixture of isotopes, which may vary from one deposit to another by as much as the shaded area in each bar, but which average to the same atomic weight.

proximity to uranium and thorium—before it was laid down in the mineral in which it is now found. The problem requires far more investigation. Calculations made from the data so far available have led the eminent geologist Arthur Holmes of the University of Edinburgh to suggest that the actual age of the earth is close to 3.3 billion years.

The mass spectrometer has many other applications to problems in geophysics and geochemistry. One of the most engaging is employed by Harold Urey and associates at the University of Chicago in determining from the fossil shells of marine animals the temperature of the oceans in which they lived millions of years ago. Urey and his associates have shown that when calcium carbonate, the material forming the shells, is crystallized slowly in water, the ratio of rare oxygen 18 to common oxygen 16 in the carbonate will depend slightly upon the temperature of the water. The effect is very small: a change

in temperature of one degree centigrade changes the ratio by only .02 per cent. Since oxygen normally contains only one part of O-18 to 500 parts of O-16, this change affects the atomic weight of oxygen by less than one 10-millionth of a unit of atomic weight. Nevertheless Urey and his colleagues, with the aid of a mass spectrometer, have found measurable isotopic differences in the oxygen extracted from fossil shells and have been able to calculate the temperatures of the oceans in which the animals lived—a valuable bit of evidence in reconstructing the early history of the earth. The technique has been perfected to the point where, by examining successive layers of an animal's shell, one can even determine seasonal variations in temperature and in some instances the time of year the animal died!

Other workers have investigated by mass spectrometry variations in the isotopic abundances of certain other elements, including boron, sulfur and

helium. Several years ago one of my students, L. T. Aldrich, and I measured the abundance of the rare isotope helium 3 in various sources of helium and found that it ranged from 5 to 2,000 parts per 100 million—a 400-fold variation! Abundant helium 4 is a by-product of the radioactive decay of uranium and thorium, while helium 3 is the decay product of unstable hydrogen 3 (tritium). Tritium, in turn, is formed in various ways, including cosmic-ray bombardment. Hence the problem of explaining the variations is extremely complex.

TO THE MEN who developed the mass spectrometer—Thomson, Aston, Dempster, the late John T. Tate of Minnesota and others—workers in many fields today are much indebted. The specialists who have put the instrument to “practical” use already include petroleum chemists, atomic fuel producers, biologists, geologists—and the list will certainly grow far longer.

The Author

ALFRED O. C. NIER, chairman of the department of physics at the University of Minnesota, built his first mass spectrometer as a graduate student at Minnesota. “I found that I had in my hands,” he says, “the best instrument then in existence for studying isotopes.” In 1935, a year before getting his doctorate, he found the extremely rare isotope of potassium, K-40. While Nier was at Harvard the following year on a National Research Council fellowship, his work attracted the attention of G. P. Baxter, the venerable authority on atomic weights. Baxter prepared special compounds for him, including uranium tetrachloride, with which Nier

determined for the first time the relative abundance of isotopes U-235 and U-238. When he returned to Minnesota in 1938, it wasn't long before he had to take up the uranium problem again. It was Nier who, by separating pure U-235 and U-238 with the mass spectrometer, made it possible to settle the question as to which was the fissionable isotope. He worked during the war with the Manhattan District and the Kellogg Corporation, builder of the U-235 gaseous diffusion plant at Oak Ridge.

Bibliography

MODERN MASS SPECTROSCOPY. Mark G. Inglish in *Advances in Electronics*, Vol. I, Academic Press, Inc., 1948.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

CHEMICAL ANALYSIS BY INFRARED

by Bryce Crawford, Jr.

Electromagnetic waves make the atoms in a compound vibrate. The frequencies to which a substance responds can tell much about its structure and about the nature of chemical bonds.

Infrared spectroscopy has grown like a mushroom in the past 10 years. Before the war it was employed by only a few chemists and physicists, using home-built or custom-built infrared spectrometers. Now the instrument is a standard commercial item supplied by a competitive industry to chemical and medical researchers all over the country.

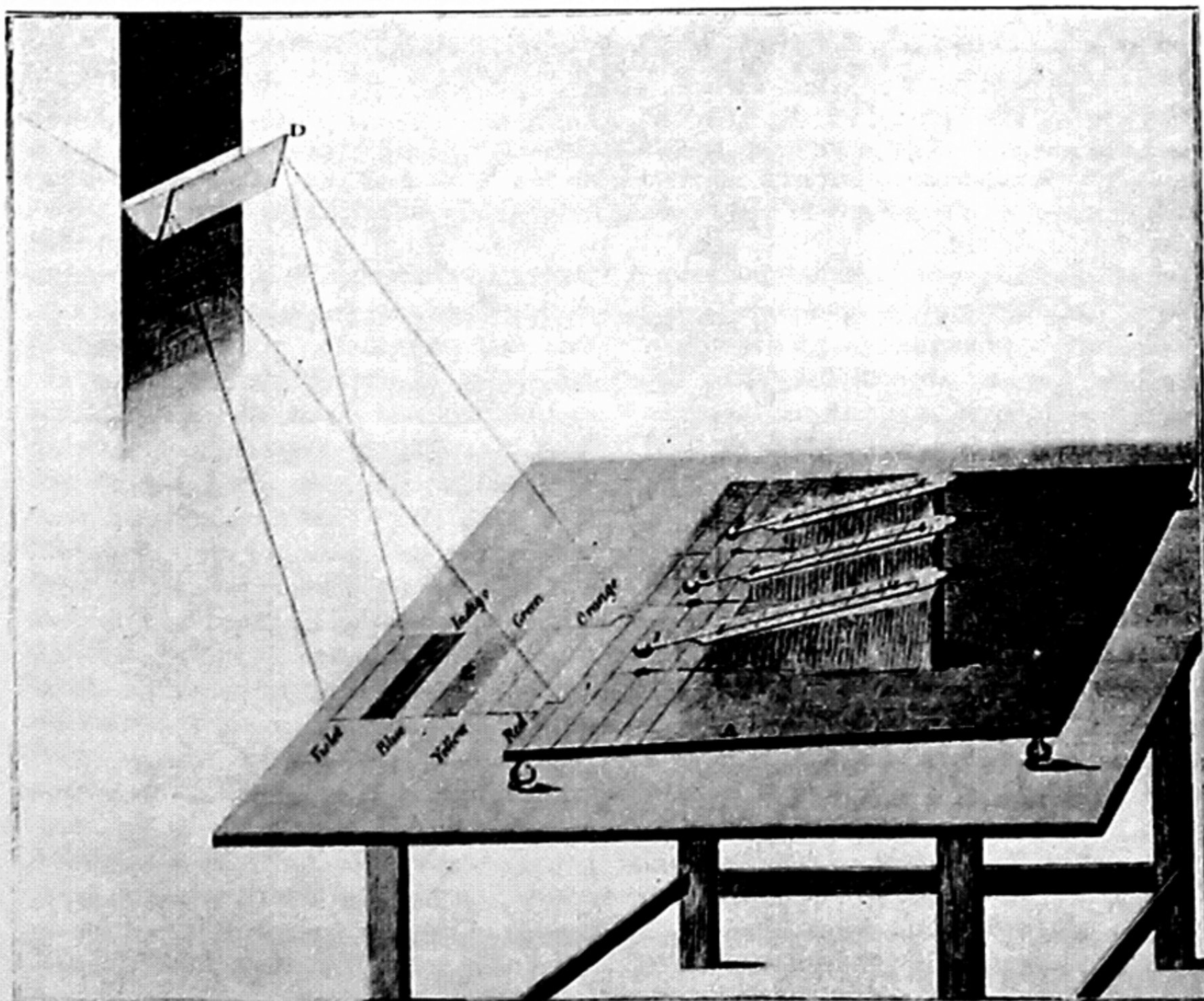
More than 1,300 commercial infrared spectrometers, each representing an investment of two to six Cadillacs, are earning their way in scientific laboratories and industrial plants.

To an old-timer in the field—"old-timer" by virtue of having studied infrared as long as, say, 15 years—this burgeoning use of infrared by his col-

leagues is most gratifying, but hardly surprising. The power of infrared as a tool for chemical characterization and analysis has been known for some 20 years. The infrared spectrum of a substance is related to its chemical structure in a uniquely convenient way, and organic chemists have been heard to say that the infrared spectrum of an organic compound is its most important physical property.

Infrared radiation itself was discovered more than 150 years ago, long before scientists had any clear understanding of radiation. Sir William Herschel, who started life as an organist at Bath and became an astronomer, made the discovery. In 1800 he reported to the Royal Society certain experiments in heat radiation. He resolved sunlight into its spectrum with a glass prism and placed a thermometer at successive positions in the spectrum. He found heat radiation not only in the visible spectrum but in the longer wavelengths beyond the red. Herschel even crudely measured the absorption of this radiation by various substances, including tap water, distilled water, sea water, gin and brandy. He could not know, could not even suspect, how revealing the absorption could be about chemical structure. Before anyone could appreciate the significance of infrared absorption, light radiation itself had to be understood. A century passed before the theory of the nature of light was worked out and the necessary techniques and instruments for infrared analysis were developed.

Sir William had found that the most intense heat radiations were outside the visible part of the spectrum. Conse-



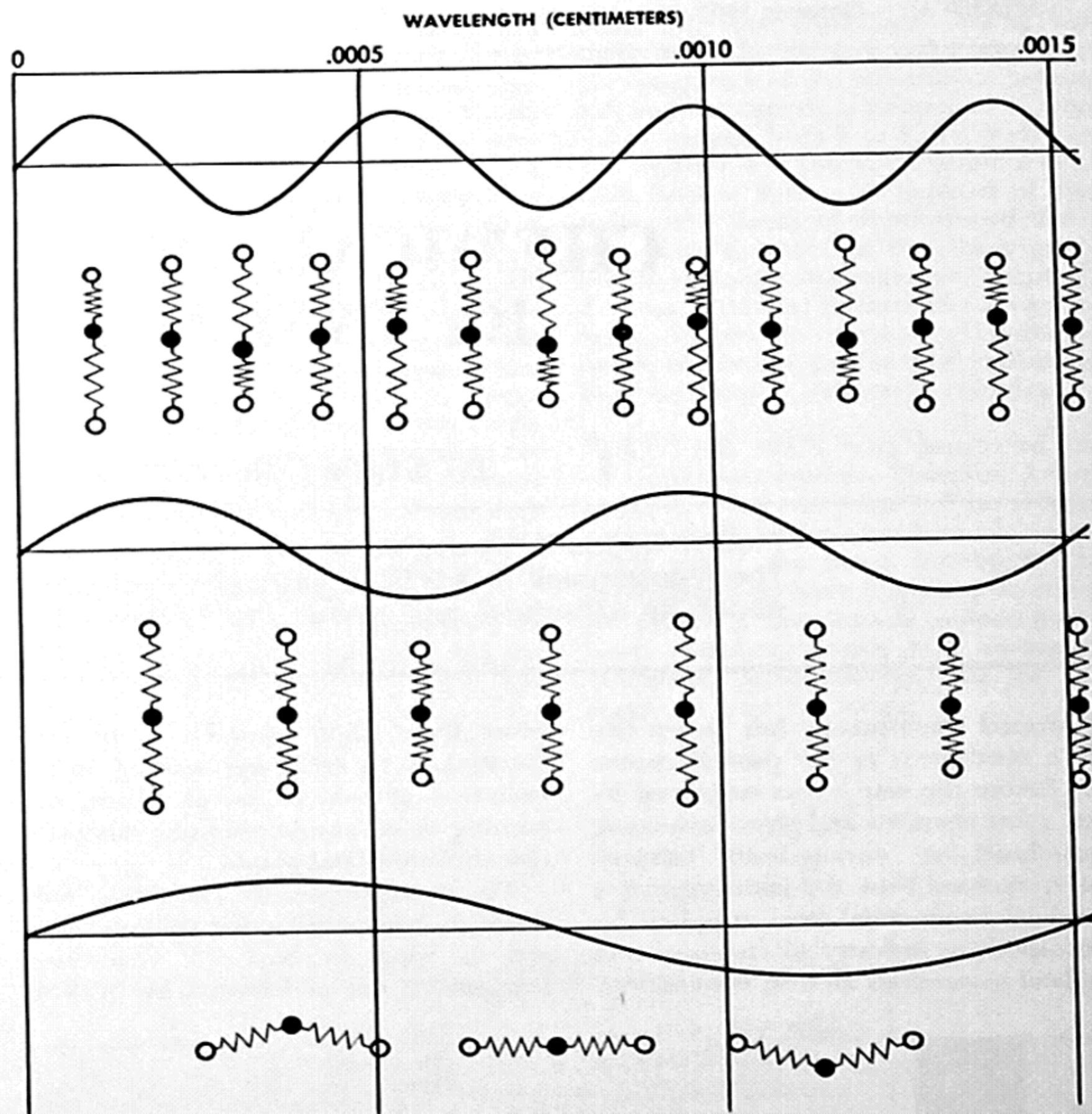
FIRST INFRARED SPECTROMETER, improvised by Sir William Herschel in 1800, was the means by which he discovered infrared radiation. Thermometers placed in successive bands of solar spectrum registered more heat outside and beyond red region than within.

quently for many years it was believed that heat and light were two quite different radiations. But some 35 years later Sir William's son, Sir John Herschel, and other investigators advanced the idea that heat and light might merely be different manifestations of the same radiation—that is, light waves of different wavelengths. Before the end of the century James Clerk Maxwell had shown theoretically and Heinrich Hertz had proved experimentally the essential identity of light, heat and other electromagnetic radiation.

The infrared region, as spectroscopists usually define it, lies between the visible and the radio portions of the spectrum—that is, the wavelengths between one thousandth of a millimeter and one millimeter. (Infrared of course is not synonymous with the common meaning of the word "heat," for, as we have seen, there are heat radiations in the visible part of the spectrum.) The region most useful to the chemist is the range of wavelengths from 2 to 20 microns (thousandths of a millimeter). This is sometimes called the vibrational region.

We shall confine ourselves to this region and to considering how absorption spectra in it give information about matter. The other infrared regions, and the physics of the radiation itself, are just as interesting: it was in the near infrared that Max Planck, by instinct and gentle nature one of the most classical of physicists, was led to the disturbing idea of the quantum, which upset the whole beautiful pattern of 19th-century physics. But we cannot cover the whole subject of infrared in one article.

The chemically useful study of infrared absorption spectra was really begun in 1903 at Cornell University. A graduate student named William W. Coblentz, under the physics professor Edward Nichols, had undertaken research on infrared absorption. After improving the available experimental techniques, he received an appointment as a research associate of the Carnegie Institution of Washington, which enabled him to set about measuring absorption spectra of pure substances. He mapped spectra for two years and in 1905 published a collection of accurate infrared absorption spectra for 131 substances. Even today, after 48 years of improvements in infrared technique (to which Coblentz himself contributed much until his retirement from the National Bureau of Standards in 1945) the monumental work that he did at Cornell still stands worthy of study. Subsequent studies



VIBRATION OF MOLECULAR BONDS accounts for selective absorption of infrared energy by various compounds. Each of three modes of vibration of carbon dioxide bonds shown here is resonant to infrared at wavelength indicated. Energy is absorbed, however, only by modes at top and bottom, which disturb geometric and electric symmetry of atom.

have confirmed more than they have corrected Coblentz's observations.

When sunlight falls upon a green leaf of a tree, the leaf absorbs the red wavelengths of the light and reflects the green wavelengths: this is what gives the leaf its color. The absorption tells something about the leaf's molecular composition. Essentially the same principle is involved in studying infrared absorption, except that the radiation is not visible; if our eyes were attuned to the infrared, we could recognize a compound by its characteristic color. Instead we measure its absorption of heat. Infrared radiation is passed through a solution of a compound, and the compound characterizes itself by the wavelengths it absorbs and those it transmits. Each compound has its own absorption spectrum.

The absorption of infrared is due to some disturbance within the molecule; this Coblentz established by observing that he got different spectra from isomers—molecules which are composed of

the same atoms but in different arrangements. Coblentz found further that certain subgroupings of atoms within molecules identified themselves by absorbing characteristic wavelengths: for example, phenyl compounds, containing the benzene ring, absorbed at 3.25 and 6.75 microns, while mustard oils, containing the thiocyanate group, absorbed at 4.78 microns. These absorptions were additive: in phenyl mustard oil Coblentz found "the characteristic vibration of the mustard oils superposed upon the vibration of the benzene nucleus." He concluded that "there is a something—call it 'particle,' 'group of atoms,' 'ion' or 'nucleus'—in common, with many of the compounds studied, which causes absorption bands that are characteristic of the great groups of organic compounds, but we do not know what that 'something' is."

With half a century's progress since Coblentz, we now have a clear idea of the "something." It lies in the bonds

between the atoms in a molecule. These bonds are written in chemical formulas as single, double and triple lines connecting the atoms. As the Danish chemist Niels Bjerrum once said, they "summarize, in a very compact form, chemistry's knowledge of the creation and destruction of compounds. Nowhere in science has a shorthand notation been developed which summarizes such an abundance of exact knowledge in so small a space." In 1914 Bjerrum showed that, if we think of the atoms as small masses and the bonds as springs holding the atoms together, we can account correctly for the vibrational behavior of molecules, as observed in their infrared spectra and in their heat capacities.

The bonds hold the atoms in position fairly tightly, but not rigidly. The response of the atoms to a light wave is much like the response of cork balls, floating on a lake, to waves on the lake. As waves move past the ball, they push

the ball alternately up and down. So a light wave moving past an atom sweeps an oscillating electric field over it, and if the atom carries an electrical charge it will be pushed first one way and then the other. Atoms in general do carry a charge, greater or lesser according to the molecule; thus in the hydrogen chloride molecule the hydrogen atom carries a small positive charge, the chlorine atom a corresponding negative one. Because of these opposite charges, the electric field of the light wave will push the two atoms in opposite directions, and will tend to set them into vibration, stretching and compressing the H-Cl bond alternately.

The bond has a natural frequency of vibration, determined by the masses of the two atoms and the restoring force of the bonds. A light wave with this frequency of oscillation will have most effect on the bond: its energy will greatly increase the natural vibrations of the

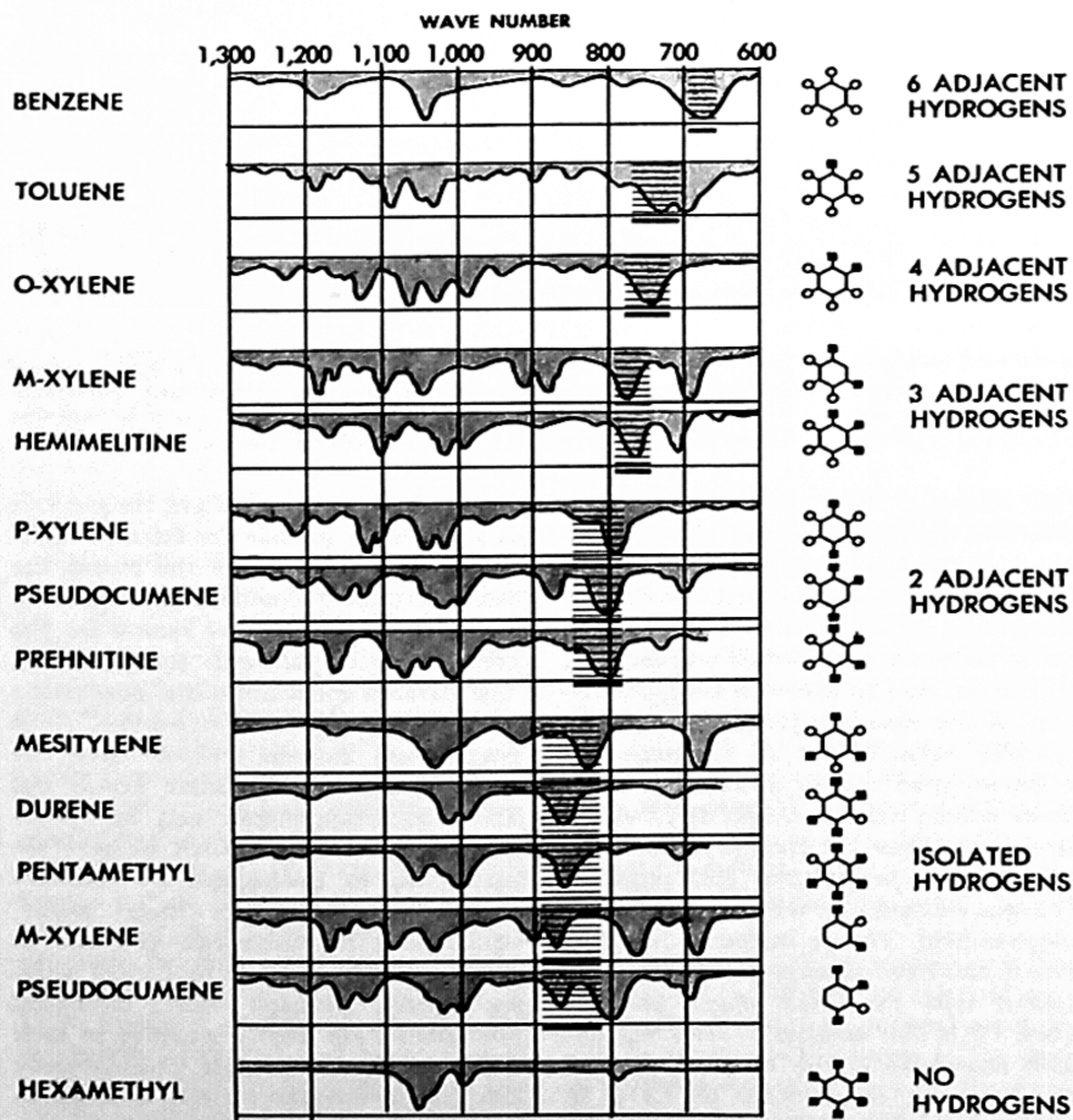
atoms. The molecule will absorb part of the energy of the light at this resonant frequency, and an absorption detector will show an absorption peak for that wavelength.

To obtain the infrared spectrum of a sample, we illuminate the sample with infrared radiation of successive wavelengths from 2.5 to 25 microns and measure with the spectrometer the amount of light transmitted by the sample at each wavelength. A modern spectrometer automatically computes the percentage of the light transmitted at each wavelength and in 10 or 15 minutes produces a curve of transmittance against wavelength, or, more commonly nowadays, transmittance against frequency. Frequency, meaning the number of waves that sweep past per second, is easily computed from the wavelength and the speed of light. The measure of frequency is called the "wave number"—it is actually the number of waves in one centimeter of light beam. In the range of infrared that we are considering the wave numbers run from 400 to 4,000 per centimeter.

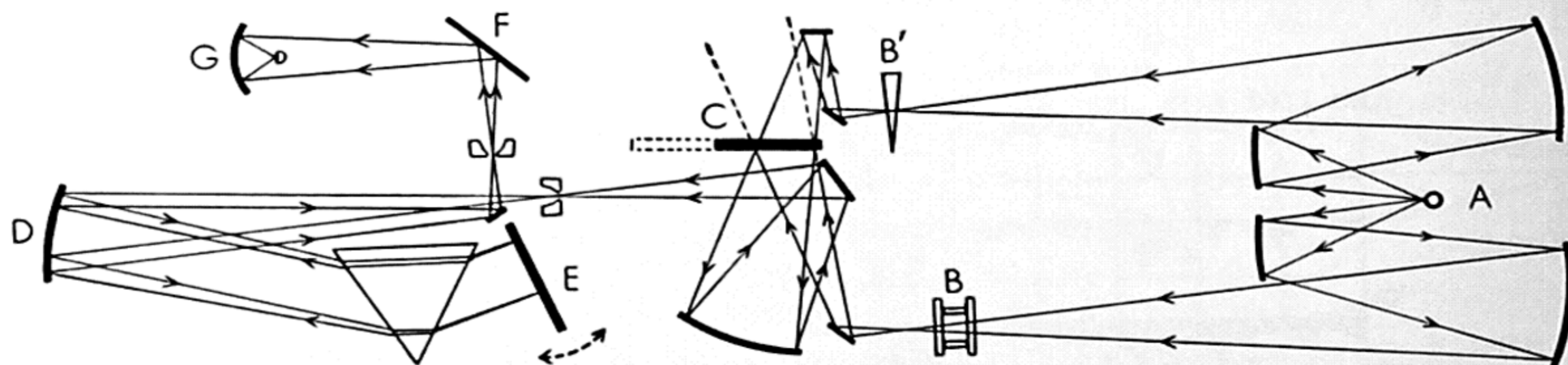
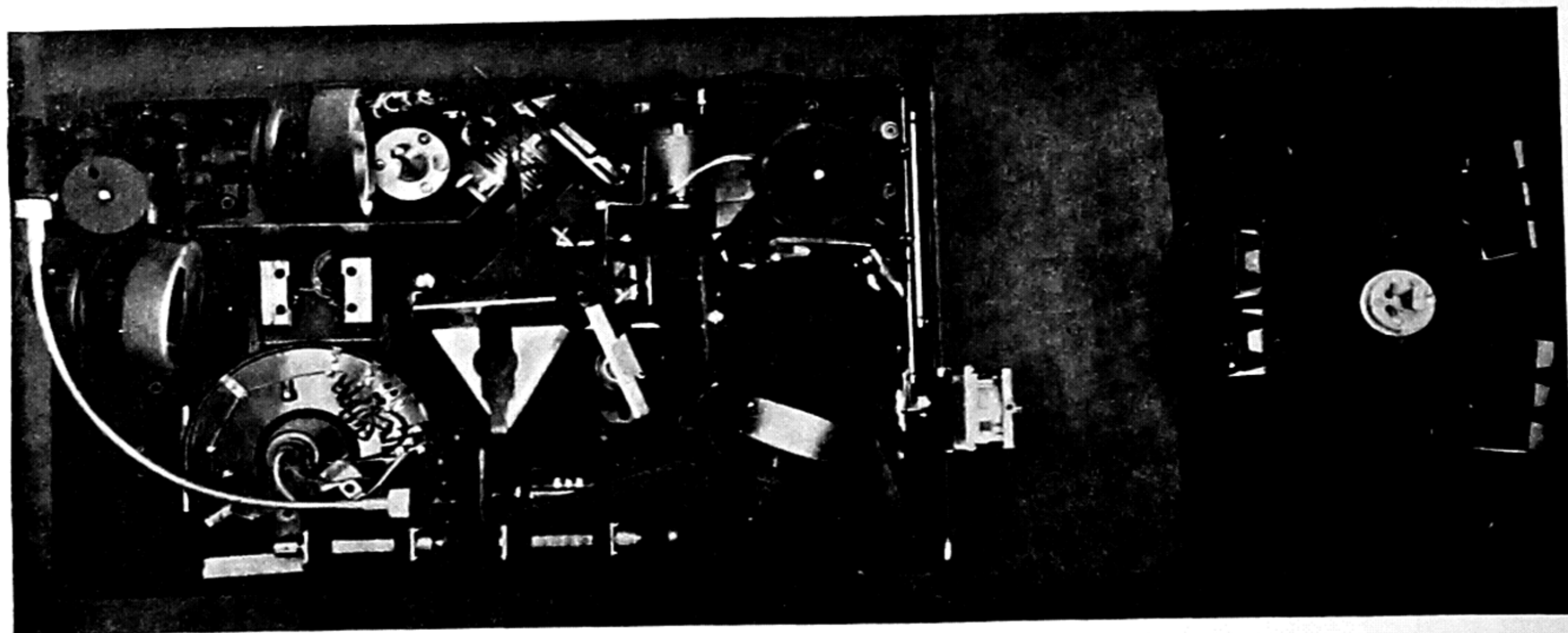
Modern atomic theory at once shows the chemical importance of the frequency of absorption. Atoms and molecules absorb light in quanta, the energy in each quantum being proportional to the frequency of the light. The absorbed energy is in fact Planck's constant h times the frequency, or hc times the wave number. Einstein enunciated the principle that, when molecules absorb light, each quantum is wholly absorbed by one molecule.

Now a molecule can safely absorb only so much energy: chemical bonds are not unbreakable—and a good thing, or there'd be no chemistry! The energy required to break the bonds is known fairly accurately; it is comparable to the energy of a quantum of ultraviolet or visible light. When a molecule absorbs such light, the bonds are either broken or profoundly altered, and the "excited" molecule is a very different entity from the original. The deduction of atomic structure from ultraviolet spectra has been compared to deducing the structure of a piano from the sounds emitted as it falls down a flight of stairs.

The absorption of a quantum in the infrared, on the other hand, is not so rough a process. At these wavelengths a quantum of energy is only about one twentieth that in the ultraviolet. Hence the radiation merely sets the bond into vibration. In infrared spectroscopy we don't push the piano down the stairs,



RELATED FAMILY of compounds, the methyl benzenes, exhibit both an over-all similarity and distinctive differences in their infrared spectra. The absorption dips which are marked in red are due to hydrogen wagging (see middle diagram on page 482). The hydrogen atoms are vibrating in and out of, in a direction perpendicular to, the page surface.



INFRARED SPECTROPHOTOMETER made by the Perkin-Elmer Corporation is shown on opposite page. Above are a photograph and diagram of its optical system. Infrared rays from a heated carbide rod (A) are split into two beams, one passing through the

sample to be analyzed (B) and the other through a wedge whose transparency can be adjusted. A rotating mirror and diaphragm (C) alternately passes the two beams, by way of a pair of mirrors, to the curved mirror (D). This reflects the beams through a prism

we just plunk the keyboard a bit, and the sounds are a bit easier to relate to the piano.

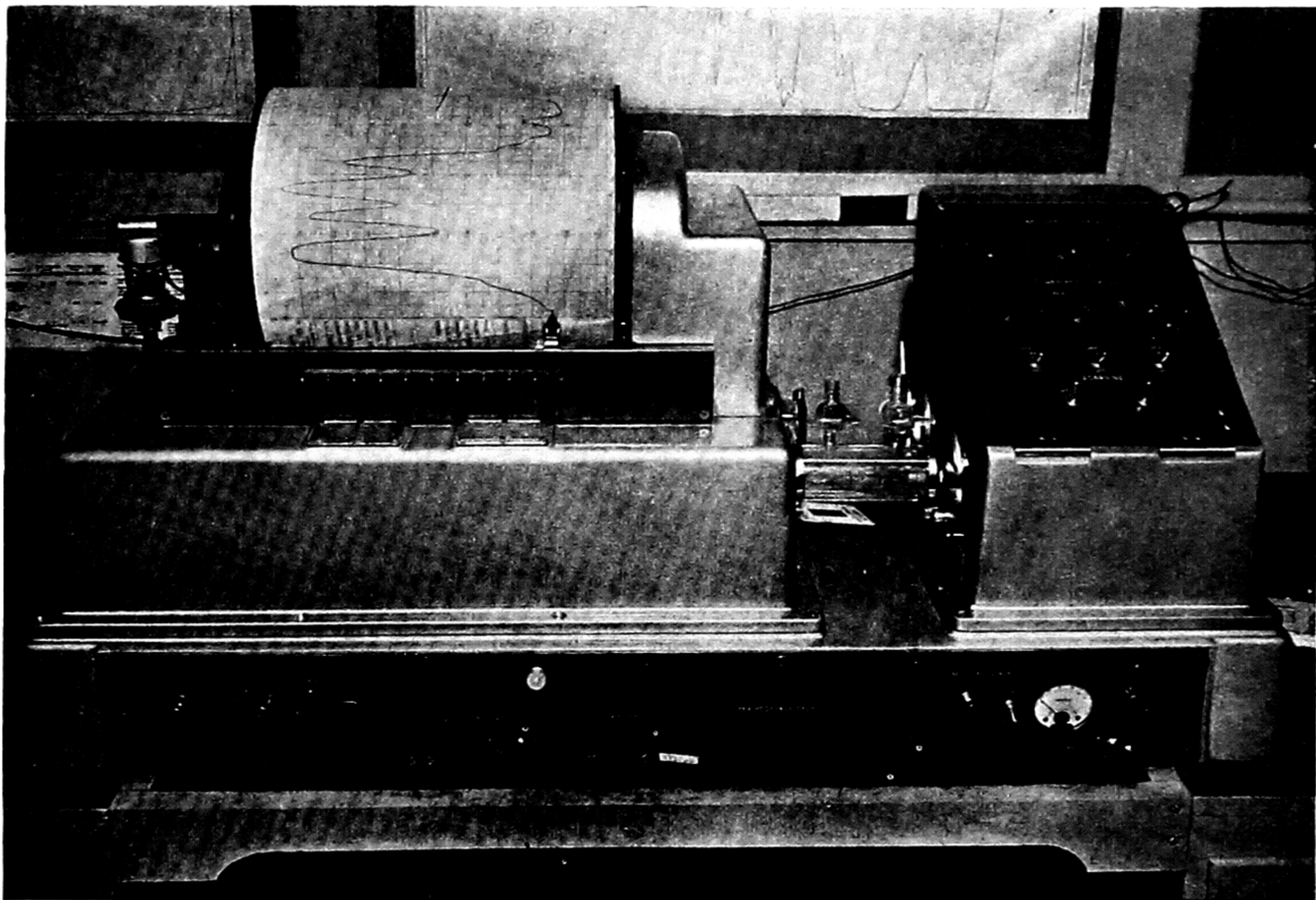
We can carry this image further. The infrared spectrum of a two-atom molecule such as HCl shows absorption not only at the single natural frequency of the bond, but also at the overtones of that frequency. A molecule containing several atoms has several bonds and several natural frequencies. Here we cannot match up the frequencies with the individual bonds. We may think of such a molecule as a mechanical system of several masses connected by springs, and the resonant frequencies will be those of such a set of coupled oscillators. When one bond is set into vibration, the rest of the molecule also is involved through the interconnecting bonds. Hence the resonant frequencies are characteristic of the *whole molecule*. They are determined by (1) the masses, which means the specific atoms involved, (2)

the spring forces, which means the bonds in the molecule, and (3) the way in which these are coupled, which means the specific geometrical arrangement of the bonds. If any of these changes, the set of resonant frequencies will change.

The infrared spectrum is simply a display of the resonant frequencies of the sample. Actually not all resonant frequencies give rise to absorption. Only those which, like the simple H-Cl vibration, cause some net change in the separation between positive and negative charges will interact with the oscillating electric field. Thus a molecule like Cl_2 , which can have no charge separation no matter how the bond length changes (one Cl is like another!), does not absorb anywhere in the infrared. One of the resonant frequencies of CO_2 , in which the two oxygen atoms move symmetrically, causes no charge displacement and hence does not appear in the infrared spectrum. But in general there

are enough active resonant frequencies in a molecule so that the infrared spectrum is characteristic of the atoms, the bonds and the geometrical arrangement.

And here we see one reason for the great value of infrared: its *specificity*. The infrared spectrum is the most nearly unique property of a substance. Even geometrical isomers, which have the same atoms and the same bonds but differ in arrangement, can be distinguished by infrared. Such isomers are very hard to distinguish by ordinary chemical methods, yet their "slight" structural differences can give rise to profound differences in biological activity. Many, perhaps most, biological phenomena are highly sensitive to such differences in compounds. Consequently medical investigators need a sensitive analytical method. The infrared spectrometer was a key tool for the late cancer researcher Konrad Dobriner at Memorial Hospital in New York City in his



which separates their wavelengths. A rocking mirror (E) sweeps the dispersed beams across the detector, by way of (D), the small plane mirror, (F), and (G). The strengths of the alternately received beams are compared at successive wavelengths. When the

detector senses a difference between the sample and the reference beams, an alternating current is generated which adjusts the movable wedge to equalize the signals. The motion of the wedge in turn is transmitted to pen that traces the absorption curve on drum.

work of unraveling the metabolism of steroids in the body.

Inanimate objects—if internal-combustion engines can be so classified—also are sometimes sensitive to isomeric differences. The difference between “knock” and “anti-knock” gasoline components is a problem nicely suited to the infrared spectrometer. Indeed, the need in the petroleum industry for a fast, reliable and convenient method of distinguishing isomeric hydrocarbons and analyzing mixtures of them has played a large part in the recent upsurge of infrared. The war gave rise to pressing demands for high-test gasoline and for synthetic-rubber intermediates; both of these involved analysis of isomeric hydrocarbons. Infrared had been used in a few industrial laboratories in the late 1930s. Under the wartime challenge it was soon shown that infrared spectra offered the rapid and accurate analytical method so badly needed. And to meet

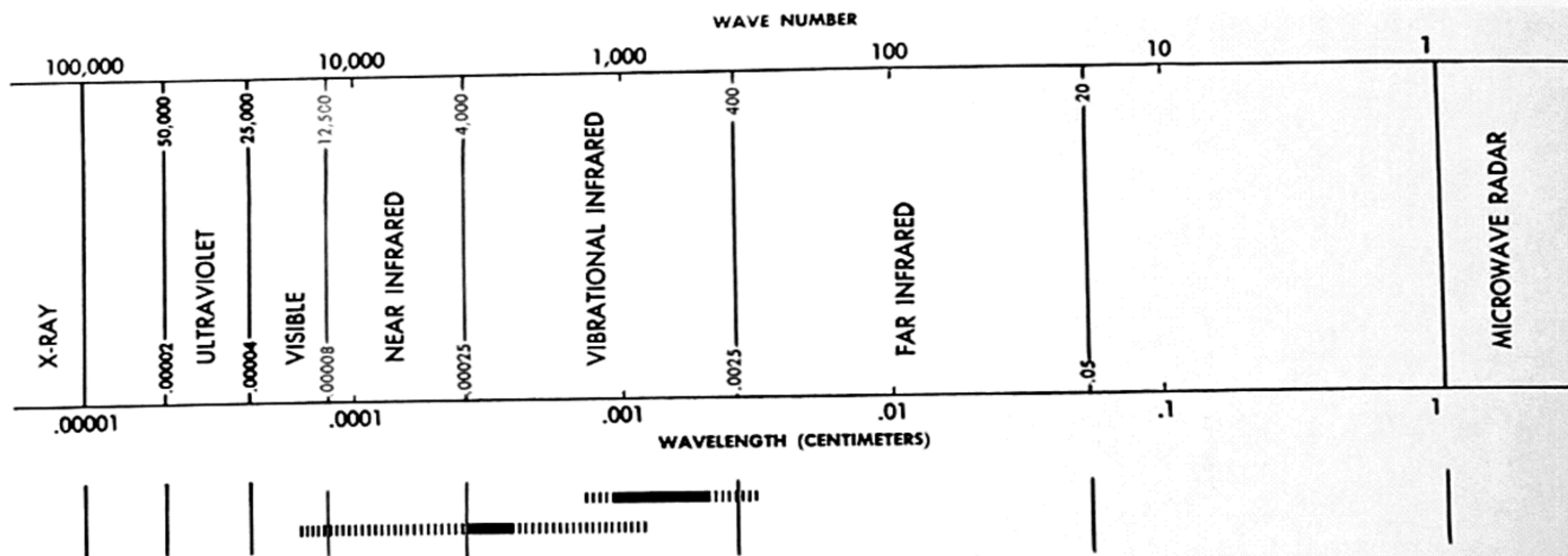
the need for instruments the first commercial infrared spectrometer appeared in 1943.

The infrared also has unusual sensitivity to atomic mass; it can distinguish not only isomers but also isotopes. Practically the only methods for analyzing the stable isotopes used in tracer work are infrared and mass spectroscopy. Which is the better method in a given case will depend on many factors; in cases calling for a fast, nondestructive analytical method applicable *in situ*, the specificity of infrared can be very useful.

Yet all this is not the whole story. The greatest advantage of infrared as a research tool is that an infrared spectrum can be interpreted in terms of the same concepts chemists use in studying chemical properties—bonds and bond groupings. In classifying compounds and thinking about them and working with them, chemists have long spoken of

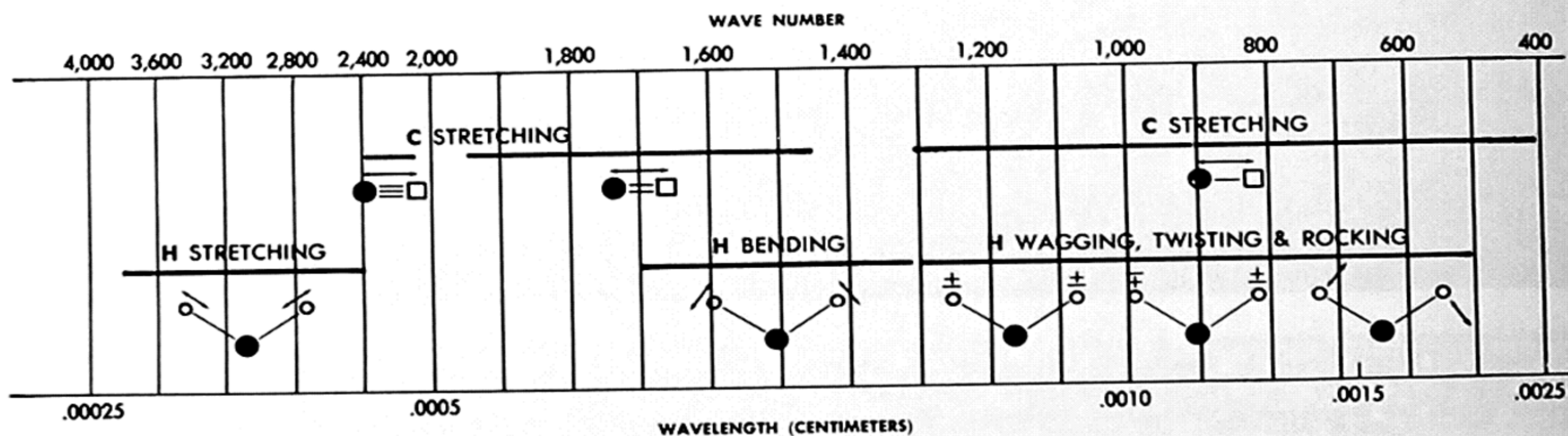
“functional groups”—of olefins, of acid chlorides and so on. The functional groups provide a broad chemical description and a clue to the specific chemical properties of a compound—and they can be related to its infrared spectrum.

Let us consider for a moment the mechanical model. When two or more high-frequency springs are coupled together tightly, the vibrations of the resulting system form a new pattern in general quite distinct from that of the uncoupled springs; but when two high-frequency springs are coupled by means of a low-frequency spring, the resulting vibration pattern will include the frequencies of the high-frequency springs with only small shifts. Now chemical bonds fall approximately into two classes: (1) “high-frequency,” which includes all multiple bonds and the single bonds involving hydrogen, and (2) “low-frequency,” which includes all other single bonds, such as C-C, C-O, C-N. In



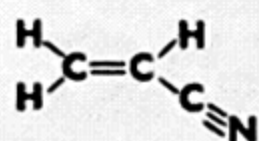
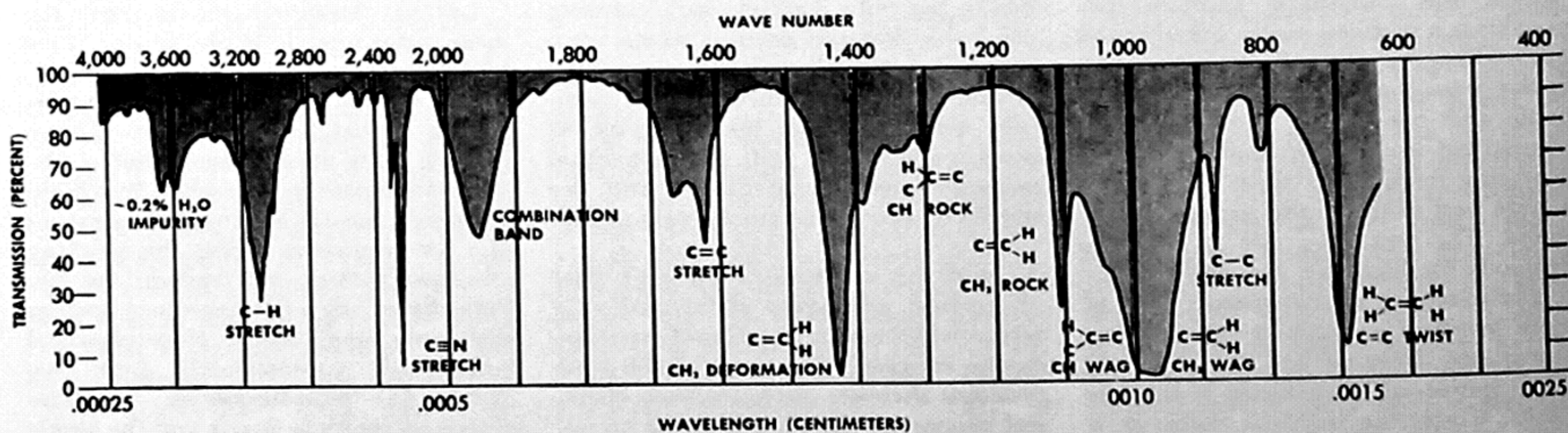
ELECTROMAGNETIC SPECTRUM is diagrammed above. Wave number means the number of wavelengths per centimeter and is proportional to frequency. Wave numbers from 400 to 4,000 have the right frequency to excite molecular vibrations. Bars below

locate concentrations of heat radiation from a black body at 80 degrees F. (*upper bar*) and 1,800 degrees (*lower*). In each bar 50 per cent of the total heat energy is in the black sections, 40 per cent in the broken section, and the rest tapers off from the ends.



TYPES OF VIBRATION that cause most common absorption bands are here located in spectrum. Carbon atoms are represented by solid black circles, hydrogen atoms by small open circles, other atoms by squares. Carbon bonds stretch at the indicated frequencies

whether running between two carbon atoms or between carbon and a different atom. Hydrogen wagging, twisting and rocking cover the 500 to 1,300 region jointly. Plus and minus signs for wagging and twisting indicate motion in and out of plane of page.



ACRYLONITRILE, whose molecular formula appears at left, gives the infrared spectrum shown above. The curve represents the amount of energy transmitted by a sample of the mate-

rial as the wavelength increases (*from left to right*). Dips indicate frequencies at which energy is strongly absorbed, showing that a natural frequency of molecular vibration has been reached. Vibrations causing dips are indicated for comparison with chart above.



DETECTOR for spectrophotometer is shown in front view (top) and side view (bottom) about 16 times actual size. Beam of radiation is focused on thermocouple, the fine gold ribbon mounted on the contacts.

studying infrared spectra we may therefore try out the idea of a "vibrational functional group"—any set of high-frequency bonds directly connected. According to this, all molecules with the group $-N=C=S$, for example, should have some frequencies in common, but they should differ from those with the $-N=C-$ or $C=S$ groups. Here indeed we find the "something" for which Coblenz was groping!

Now we find a piece of good fortune—one of the rare presents Nature bestows on investigators. Chemical functional groups and vibrational functional groups run parallel. Therefore certain characteristic absorption bands in an infrared spectrum give direct and strong evidence on the chemical nature of the sample. It is not really conclusive evidence, for the functional group idea is an approximation, and the parallelism of chemical and vibrational functional groups is not too strict. The infrared spectrum is not a magic crystal ball in which one reads the structural formula of an unknown sample. But its clues, when wisely used, can shorten by weeks the time required to complete the job. And the direct applicability of the chemical-bond concept and the functional-group concept means that the chemist can understand infrared spectra without having to learn a new language. He *does* have to learn a new dialect of his chemical language—and incautious chemists who have overlooked this have made some serious blunders. But the dialect can be picked up relatively quickly.

The knowledge of molecular structure required to make use of infrared spectroscopy had been achieved by about 1930 or 1935. Why, then, did the blossoming of chemical infrared start only in 1943? Some pioneer applications of infrared *had* been made in the chemical industry as early as 1936. But the technique is difficult. The cost of the modern commercial instruments and their need of maintenance still remind the user that the infrared spectrometer is doing a basically harder job than its ultraviolet counterpart. It has been well said that "no other spectrometric region of the electromagnetic range is so beset with experimental difficulties." Basically these arise from the very fact that infrared radiation is resonant with atomic vibrations. We have seen the advantages stemming from this; let us glance at the price we must pay.

As a source of infrared, we must use vibrating atoms, and the only practical way to set atoms vibrating is through heat. To get a reasonable intensity of

infrared, we must heat our source to a high temperature; a common source is an electrically heated rod of carborundum at 2,200 degrees Fahrenheit. Refractory materials able to withstand such temperatures are brittle, which can be a nuisance! Moreover, the laws of physics worked out by Planck show that most of the energy goes into the near-infrared and even the visible region, wasting our power and filling our instruments with a lot of unwanted wavelengths which must be filtered out or shunted off.

The same basic trouble arises in the measurement of infrared intensity—detection, as it is appropriately called. The infrared quanta do not disturb molecules enough even to affect a photographic plate! All we can measure is a slight rise in the temperature of the absorber, and we measure it today as Herschel did—by the effect on a blackened thermometer. The "thermometer," however, is a very sensitive thermocouple.

We must find prisms to disperse the infrared spectrum and optics to focus it, windows to let the radiation through the cells holding our samples and solvents in which to dissolve materials we want to study. Since all atoms can vibrate, it is not easy to find substances transparent to the infrared. Glass and water are completely opaque to it; ordinary organic solvents have so many absorption bands themselves that they obscure the spectrum of the substance dissolved in them. So we use mirrors instead of lenses, make our prisms and cell windows of rock salt, and think hard about the choice of solvents.

For infrared, the wartime need was a blessing in disguise—a very perfect disguise, to quote Mark Twain—because urgent necessity stimulated great improvements in instrumentation. The availability of better instruments today is a great blessing not only to analytical chemists and workers in the chemical industry but also to infrared spectroscopists interested in fundamental chemistry. Infrared spectroscopy has helped us win our present understanding of molecular structure, notably the geometry and dynamics summarized in the ball-and-spring model. Nowadays more people than ever before are at work on this fundamental use of infrared, studying the nature of those springs and the distribution of electronic charges in the bonds. Without depreciating the more widespread use of infrared for the analysis of compounds, we may well feel that in the long run the fundamental use will be more exciting. For the nature of the chemical bond is the problem at the heart of all chemistry.

The Author

BRYCE CRAWFORD, JR. is chairman of the department of physical chemistry at the University of Minnesota. He was born in New Orleans in 1914 and was educated at Stanford University, where he took his Ph.D in 1937. He spent two years at Harvard University as a National Research Fellow and then taught chemistry at Yale University for a year. He has been at Minnesota since 1940. During the war Crawford did research on rocket propellants for the National Defense Research Committee and was awarded the Presidential Certificate of Merit. He has held a Guggenheim fellowship and a Fulbright grant. His research specialty is molecular structure, which he studies largely by means of

the techniques described in his article. In leisure hours he devotes himself to a scholarly interest in the treatises of Dr. Watson (but has been too busy to contribute to the *Baker Street Journal*), to model railroading and to the care and feeding of two sons.

Bibliography

- INFRARED SPECTROSCOPY. R. Bowling Barnes, Robert C. Gore, Verner Liddel, Van Zandt Williams. Reinhold Publishing Corporation, 1944.
- INFRARED AND RAMAN SPECTRA. Gordon B. B. M. Sutherland. Methuen & Company, 1935.
- INFRARED AND RAMAN SPECTRA OF POLYATOMIC MOLECULES. Gerhard Herzberg. D. Van Nostrand Company, Inc., 1945.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

TITANIUM

by George A. W. Boehm

Nearly as strong as steel, but only half as heavy, it will soon join iron and aluminum, metallurgists believe, as one of the three most important metals.

TITANIUM is a rather common element that has been known for more than 150 years. It is the ninth most abundant element in the earth's crust—more than 20 times as common as carbon, more plentiful than lead, copper, tin and half a dozen other metals combined, outranked among the metals only by aluminum, iron and magnesium. Yet as a useful metal titanium has been discovered only within the past decade.

It now seems to be on the threshold of a brilliant career. Pure titanium is a silvery metal which looks and behaves like stainless steel, but is little more than half as heavy. It is much lighter than iron, much stronger than aluminum, almost as rustproof as platinum. Although it is not preëminent in any single quality, it combines the best features of many metals; like an Olympic decathlon champion it is outstanding in all-round ability. The metal has certain drawbacks: it will never be as cheap as steel, and its refining, casting and fabrication present difficulties. But its combination of light weight and strength and its exceptional resistance to corrosion make it ideal for a great variety of uses, from pen points to engines and aircraft frames. Metallurgists and engineers believe that titanium will soon become one of the world's most important metals.

Titanium is a "new" metal because it has only recently been refined sufficiently to reveal its properties. It is still being produced only on an experimental basis, at the rate of a few hundred pounds per day. But the metal is under investigation in some 100 laboratories, and the investigators are almost uniformly enthusiastic about its possibilities. Their studies of course include the alloying of titanium with other metals. Since pure titanium is comparable in strength and wearing qualities with stainless steel, its alloys should be truly remarkable.

Titanium will not replace iron or aluminum, but it will bridge the gap between them. The three metals will form a complementary trio. Iron and steels are strong and cheap, but not light. Aluminum alloys are light and cheap, but not particularly strong. Titanium alloys will be light and strong, but not cheap.

Thus titanium will be reserved for the many applications in which its exceptional combination of qualities outweighs price considerations.

TITANIUM was discovered in 1791 by William Gregor, an English clergyman and amateur chemist, who identified compounds of the metal in the black sands of Cornwall. A few years later the German chemist Martin Heinrich Klaproth, now famous as the discoverer of uranium, rediscovered Gregor's element in an ore called rutile. He named it titanium after the Titans, who in Greek mythology were the sons of the Earth.

The richest ore of titanium is a black oxide mineral called ilmenite, after the Ilmen Mountains in Russia. Ilmenite is also rich in iron. For hundreds of years this ore had been mined in the Ilmen Mountains for its iron, the titanium being discarded as an unwanted impurity. But at about the end of the 19th century the first use was found for the element. The aluminum and steel industries began to employ extremely small amounts of rather impure titanium as an alloying ingredient. In this form titanium was so brittle that it bore a closer resemblance to rock than to metal. A fraction of one per cent of titanium was found, however, to refine the crystal structure of some steels and aluminum alloys, thus endowing the metals with greater toughness.

A more important use, which gave birth to a new industry, was discovered in 1908. Dr. A. J. Rossi of the Titanium Alloy Manufacturing Company prepared some white titanium dioxide (TiO_2) and mixed it with vegetable oil to make a white paint. He found that the compound made an excellent white pigment. In 1926 a process was developed for manufacturing a pure grade of titanium dioxide, and industry began to take a great interest in the material. As a pigment it is outstanding for its opaqueness, or hiding power. One pound of titanium dioxide will cover a larger area than five pounds of lithopone, a very good pigment made of zinc and barium. Since the war titanium dioxide

has increasingly replaced white lead, which is now hard to obtain.

Titanium dioxide is also useful in conserving paper. When impregnated with this dense pigment, the thinnest grades of paper are opaque and can be used for Bibles and air-mail stationery. Large amounts of it are rubbed into rayon to dull its natural gloss. Today the production of titanium dioxide in the U. S. is an estimated 225,000 tons a year. The material is so important that it has recently been the subject of an antitrust order by the courts.

The infant titanium metal industry as yet is only a tail on the kite of the pigment industry. Chemists are still struggling with the problem of refining the metal. The mining of titanium offers no particular difficulties. Rich deposits of ilmenite are already being worked in New York, Virginia and North Carolina, and certain black beach sands in Florida and Oregon are excellent sources of the element. Perhaps the most promising deposits on the North American Continent are those in the Province of Quebec.

In theory, the refining of titanium is simple enough. As worked out by Dr. Wilhelm Kroll, a metallurgist from Luxembourg, the process involves separating titanium from its accompanying iron in impure oxide form, and treating it with chlorine to form titanium tetrachloride, a volatile liquid. This compound is then mixed in a closed iron chamber with metallic magnesium, which because of its greater chemical activity preëmpts the chlorine, reducing the titanium to the metallic state. The final step is to get rid of the magnesium chloride and any excess magnesium metal, either by evaporating them by heating in a vacuum or by washing the mixture with hydrochloric acid, which dissolves the magnesium. The process yields titanium in a gray sponge form, which is then melted and cast into ingots.

In practice, however, titanium refining is attended by problems at every step. The metal is hard to handle on heating, for at high temperatures it becomes extremely active. It greedily absorbs oxygen and nitrogen from the air. And as little as one per cent of these gases utter-

ly destroys the ductility of titanium, rendering it as brittle as pumice. So at every stage of processing, titanium must be protected from air, either by a blanket of inert gas (*e.g.*, helium or argon) or by a vacuum. Moreover, the step in which titanium tetrachloride is reduced with magnesium presents another difficulty. If the temperature is not carefully controlled and kept below 1,600 degrees Fahrenheit, the titanium attacks and dissolves the walls of the iron reaction chamber. Control of the temperature is complicated by the fact that the reaction gives off a tremendous amount of heat, and titanium is a relatively poor heat conductor; thus it does not appear possible to draw off the heat fast enough unless small reaction chambers are employed. This will probably restrict the production of titanium to 250-pound batches, unless a method radically different from the Kroll process is devised.

THE most exasperating difficulty of all is that of melting the purified metal sponge to cast it. One metallurgist who has been struggling with titanium asserts that the problem of producing the pure metal at low cost is just that of melting it cheaply and in large quantities. Titanium melts at 3,140 degrees F. It is melted in an electric furnace, and of course must be handled in a vacuum or an atmosphere of inert gas. However, in its molten state titanium not only absorbs nitrogen and oxygen but dissolves crucibles and furnace linings made of the common resistant materials. Glassy oxide furnace linings such as aluminum oxide, which are used to hold other molten metals, are ruinous to titanium. Even the most stable of these materials yield their oxygen to the avid metal. Current research on this problem points to dense carbon crucibles as the most satisfactory. Small quantities of carbon in themselves do titanium no great harm. But since the total amount of impurities that can be added to titanium without seriously impairing its ductility is strictly limited, another crucible material must be used when carbon is not to be a final constituent of the titanium alloy.

The Battelle Memorial Institute of Columbus, Ohio, has developed a technique for melting titanium without contaminating it. The crucible is a copper hemisphere. The key to the success of this method is a bath of cold water surrounding the crucible. Metallurgists at Battelle have found that molten titanium does not wet cool copper; it stands aloof from the surface of the copper like water on a duck's back. This method, however, has certain difficulties, and some metallurgists doubt that its cost can be reduced sufficiently to make it generally applicable.

Still another possible crucible material is thorium oxide. Like other oxides,

the material contaminates the molten titanium, but in this case the contamination may prove to be welcome, for preliminary tests indicate that a trace of thorium greatly reduces titanium's unseemly appetite for oxygen.

Once it has been melted and cast into ingots, titanium need no longer be kept in a vacuum. It can without difficulty be forged, hot-rolled, cold-worked, drawn into wire, extruded as tubing, or machined. When it is kept above about 1,300 degrees F., however, titanium must be sheathed between sheets of steel to keep it from absorbing oxygen and nitrogen. In general it is handled like other metals; indeed, its mechanical properties are so similar to those of stainless steel that it can be fabricated with the same equipment with only minor adjustments. After prolonged heating, titanium acquires an extremely hard, black surface scale which must be removed by grinding with carbide tools, but otherwise it is no more difficult to drill and cut than stainless steel. While no way has been found to braze or solder titanium or to weld it to metals other than itself, the metal has excellent shear strength. Thus titanium rivets are strong, and rivet holes in sheet titanium also hold their shape under great stress.

But it is on titanium's alloys that metallurgists pin their hopes. There is good reason to believe that these alloys will be as far superior to pure titanium as special steels are to iron. The metal is presently being blended with virtually every other element in the periodic table. The results so far are fragmentary, but one positive statement can be made: alloys for structural use will all be titanium-rich, containing not more than about 10 per cent of other elements. This limit is necessary because the ductility of titanium is so sharply reduced by any other substance.

THE first large-scale use of titanium will probably be in the aircraft industry. Because the light metals aluminum and magnesium lose their strength at temperatures much above 350 degrees F., aircraft designers must use more and more steel in modern high-speed planes. Yet every pound of extra weight adds an estimated \$25 to \$250 to the fuel cost during the life of the plane. Wherever titanium replaces steel, it can cut the weight by 40 per cent. On this basis titanium in planes would be economical at two or three times its present cost.

Titanium is strong and elastic up to about 800 degrees F. Above that point it weakens rapidly, and prolonged exposure to temperatures greater than 1,000 degrees causes the metal to soak up embrittling oxygen and nitrogen. It should be possible, however, to alloy titanium so that it will behave better at high temperatures. Some success in this direction has already been achieved.

Ironically it has been found that titanium's performance at high temperatures is greatly enhanced by alloying it with small amounts of aluminum. Titanium has a low coefficient of thermal expansion—about half that of most steels. It also retains its strength and flexibility at subzero temperatures, an important consideration in high-flying airplanes.

Titanium has already been suggested for a wide variety of other applications in industry and consumer goods. Its resistance to expansion and contraction makes it useful for precision parts in instruments; it will stay within prescribed tolerances over a wide temperature range. Because it can be surface-hardened to a remarkable degree, it is an ideal material for spindles in cotton mills, where the continuous rubbing of thread rapidly erodes most metals except hardened steel, which is heavy and consumes a great deal of power. Titanium is impervious to the most corrosive foods—cider, vinegar, onion juice, pineapple juice, lard, tea, coffee, grapefruit juice—and so it is a promising candidate for pipes and tanks in food processing plants. Many of the strongest acids and alkalis, even when hot and concentrated, have no effect on titanium. There is a strong possibility that its remarkable inertness will make titanium



ILMENITE QUARRY of the National Lead Company in the Adiron-

suitable as a replacement for bone and cartilage in medicine. Several surgeons are now testing titanium's biological acceptability by implanting the metal in living bodies. If muscular tissue adheres to the metal and no irritation is set up, titanium may replace tantalum and silver for this purpose. Titanium's light weight should be a great asset in this application.

Some enthusiastic naval officers envision an entire battleship built of titanium. The Navy is perhaps titanium's most ardent admirer, for the very good reason that titanium has a truly amazing resistance to salt-water corrosion. The results of one experiment show why the Navy is excited. A sheet of titanium was hung half-submerged in the ocean and buffeted by a thin, fast jet of sea water. After two months of this rough treatment, the titanium was removed. The experimenters had neglected to mark the piece of metal, however, and the titanium was so clean that they were unable to tell on which side the jet had played. All other metals except platinum and Hastelloy C disintegrate during this test, and the Navy is not anxious to rely on Hastelloy C, which is composed largely of the critical metals nickel and chromium.

Alloy research on titanium is hampered by uncertainty because of the

slight impurities present in the "commercially pure" metal. Variations of the Kroll process followed by careful melting yield 99.5 per cent pure titanium. But the metal's properties are so greatly altered by the little remaining iron, manganese, oxygen, nitrogen and hydrogen that metallurgists must turn to an even purer form in order to isolate the effects of added alloying ingredients. For research purposes they have refined it to a form known as iodide titanium which is more than 99.99 per cent pure. It is made by plating titanium from titanium iodide vapor on an incandescent tungsten wire. Iodide titanium varies in price from about \$200 to roughly \$3,000 a pound, depending on the difficulties encountered in a month's processing. The iodide process is important for metallurgical research, but as an industrial production method it is out of the question.

TITANIUM'S chemical activity is its greatest weakness and at the same time one of its most attractive properties. Its suicidal affinity for gases when hot, and for virtually everything when molten, makes titanium inherently difficult and expensive to process. On the other hand, its inertness when cool will enable titanium to replace aluminum, magnesium and even stainless steel for

many purposes.

The eccentric behavior of titanium can be explained rather simply. It is actually an active chemical element, in a class with aluminum and magnesium. This accounts for the activity observed when the metal is hot or molten. When cooled and exposed to air, however, titanium acquires a thin, invisible coating of oxide and nitride which serves as an inert, impermeable chemical armor. Aluminum and magnesium are protected by passive coatings of the same type. Since these metals are purified by electrolysis and in commercial practice are never in the molten state, their natural chemical activity poses no problem. This indicates that any improved process which may be devised for purifying titanium will be electrolytic.

E. I. du Pont de Nemours & Company, the first industrial organization to produce ductile titanium, sells the metal in sponge form for \$5 a pound in 100-pound lots. Hundred-pound ingots cost \$7.50 a pound. Even when mass production gets under way, probably in the early 1950s, the price is not expected to fall below \$1 a pound for many years. Kroll process titanium will cost somewhere in the range of 60 cents to \$1.25, when the metal reaches full commercial stature. A standard stainless steel, with 18 per cent chromium and eight per cent



Black Mountain is one of the chief present sources of titanium. The ore is plentiful in the U. S. Ilmenite

also yields iron, which makes its mining doubly fruitful. This mine also produces the rich iron ore magnetite.

■ TITANIUM
 □ STAINLESS STEEL
 ■ 75-S ALUMINUM ALLOY

WEIGHT PER CUBIC INCH

STIFFNESS (OR MODULUS)

MELTING POINT

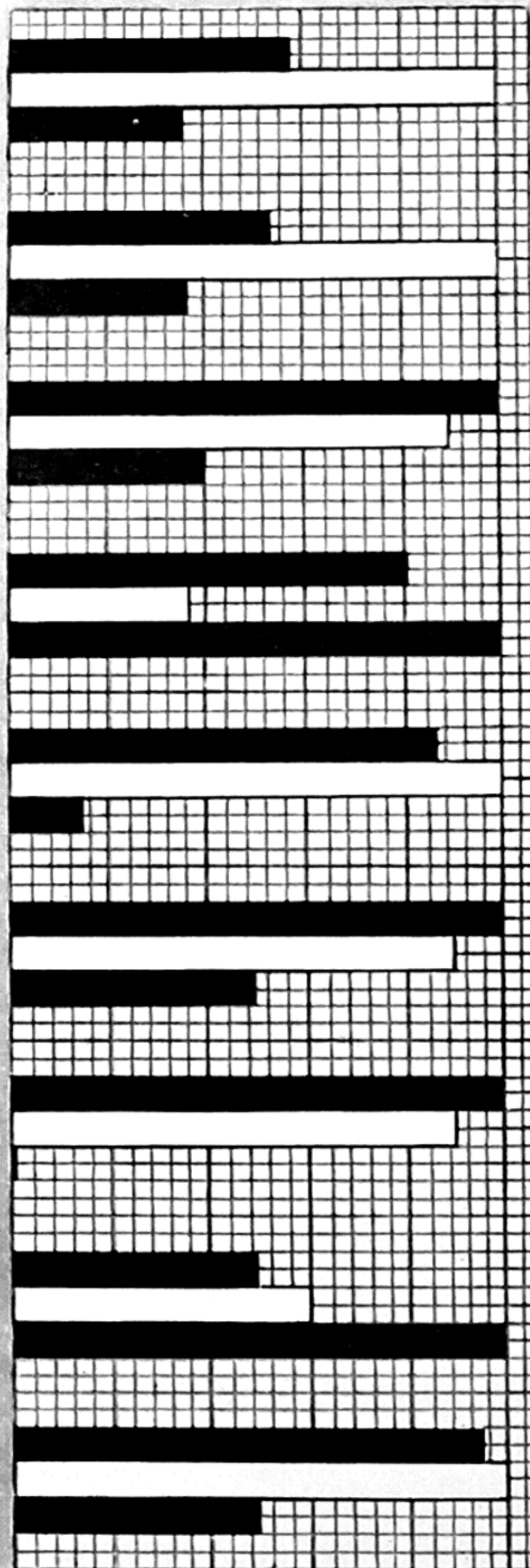
STRENGTH PER UNIT OF WEIGHT

STRENGTH PER UNIT OF WEIGHT
AT 800 DEGREES F.RESISTANCE TO SEA WATER AND
ATMOSPHERIC CORROSION

2000 DEGREES F. FLAME TEST

MACHINABILITY

WELDABILITY



TITANIUM'S PROPERTIES, here compared with those of stainless steel and an aluminum alloy, make it an almost ideal metal for most purposes. Its usefulness will be limited only by cost and some difficulties in production.

nickel, sells for about 35 cents a pound, but on the other hand special steels bring as much as \$3. Titanium's weight advantage is a vital factor in the price picture, for \$1 titanium will be on a par with 60-cent steel. And the fact that titanium and iron go hand in hand in nature will make it economical to exploit low-grade ores to the advantage of both metals.

Titanium today is in the position held by aluminum at the turn of the century. Although metallurgists and engineers foresee a brilliant future for this new and exciting metal, they can no more foretell its ultimate applications than men 50 years ago could predict the use of aluminum in automobile motors and aircraft. It is reasonable to assume, however, that in this age of close-knit technology titanium will grow toward commercial maturity at a far more rapid pace than aluminum. The rate of development of the titanium industry, compared with that of the aluminum industry, probably will be proportional to the development of atomic energy compared with the sluggish early progress in electricity.

Titanium is only the first of several chemical elements scheduled to make their debuts as metals within the next few years. Another is zirconium, which is now ready for commercial development. It is heavier than titanium, but even more resistant to corrosion and high temperatures. Boron, vanadium, and silicon, the latter the earth's most plentiful element, also show some promise as structural materials.

All of the new metals have one thing in common: abundance. Their advent as useful metals represents something new in metallurgy. Many resources of the convenient, easy-to-process metals, such as lead, tin, zinc and copper, are running low. Faced with shortages that were brought to a head by the lavish wartime expenditure of materials, we are forced to turn to unfamiliar metals, which are fortunately plentiful.

Much of the development of titanium and the other new metallurgical candidates is due to the foresight of the U. S. Bureau of Mines. In 1935 the Bureau's Metallurgical Division under Dr. Reginald S. Dean embarked on a program with two objectives: to develop methods of processing low-grade ores, and to evaluate the worth of abundant but little-known metals. This research has already accumulated dividends that we may be able to begin collecting in the near future.

Among the new metals titanium has now arrived, and others will swiftly follow in its footsteps. It is entirely possible that in years to come, this new race of metals will virtually dispossess all of today's common metals with the exception of aluminum, iron, and magnesium.

The Author

GEORGE A. W. BOEHM is a writer on scientific subjects.

Bibliography

TITANIUM AND ZIRCONIUM: METALS OF INDUSTRIAL PROMISE. William Waggaman and Edwin A. Gee in *Chemical and Engineering News*, Vol. 26, No. 6, pages 377-381; February 9, 1948.
TITANIUM. Bruce W. Gonfer in *Journal of Metals*, Section I, pages 6-9; January, 1949.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

ZIRCONIUM

by Stephen M. Shelton

Extraordinarily resistant to heat, corrosion and hard radiation, yet malleable and ductile, the metal meets some of the stern engineering requirements of nuclear reactors and jet engines.

ZIRCONIUM has recently come to the attention of metallurgists and engineers as a structural metal having tremendous potentialities. It has been referred to as the "Cinderella" metal. Because it is highly resistant to heat and radiations, it is a promising structural material for nuclear reactors and may play an important part in harnessing atomic power. It is also of great interest to engineers seeking a metal that can withstand the terrifically high temperatures in the combustion chambers of jet engines and super-rockets. In addition, zirconium is remarkably resistant to corrosion—even more resistant than tantalum, which has heretofore been considered the most impervious metal. Zirconium metal has other valuable properties as well. Science and industry are beginning to realize that the new wonder metal is almost certain to play a major role in this country's rapidly moving technological development.

Zirconium used to be considered in the class of "rare" elements, but actually it is fairly abundant in the earth's crust—more abundant than nickel, copper, lead, zinc or some other familiar metals. Its chief mineral, called zircon, occurs in sands. The largest present sources are beach sands in Australia, India and Brazil, but there are commercial quantities in this country also, in the beach sands of Florida and Oregon and the vast placer sands of Idaho. The main supply problem with regard to zirconium is not the mineral but production of the refined metal, which presents some technical difficulties. These are being solved, however, and substantial production of purified zirconium metal is now under way in the U. S.

This element with the new-found qualities and the strange name, which comes originally from a Persian word meaning gold-colored, is not precisely a new discovery. The mineral zircon (zir-

conium silicate) was known to students of gems as long ago as the early part of the Christian era. They called it hyacinth, from the Greek word for a flower or a gem. The particular mineral to which they gave this name was a reddish-orange variety of zircon found in the gem-bearing gravels of Ceylon. The element zirconium itself was discovered in 1789 by the German chemist Martin Heinrich Klaproth, who is now famous also as the discoverer of uranium, titanium and cerium. He identified zirconium while analyzing a precious stone from Ceylon. In 1824 the Swedish chemist Jöns Jakob Berzelius succeeded in extracting from zircon some zirconium metal, though in an impure state.

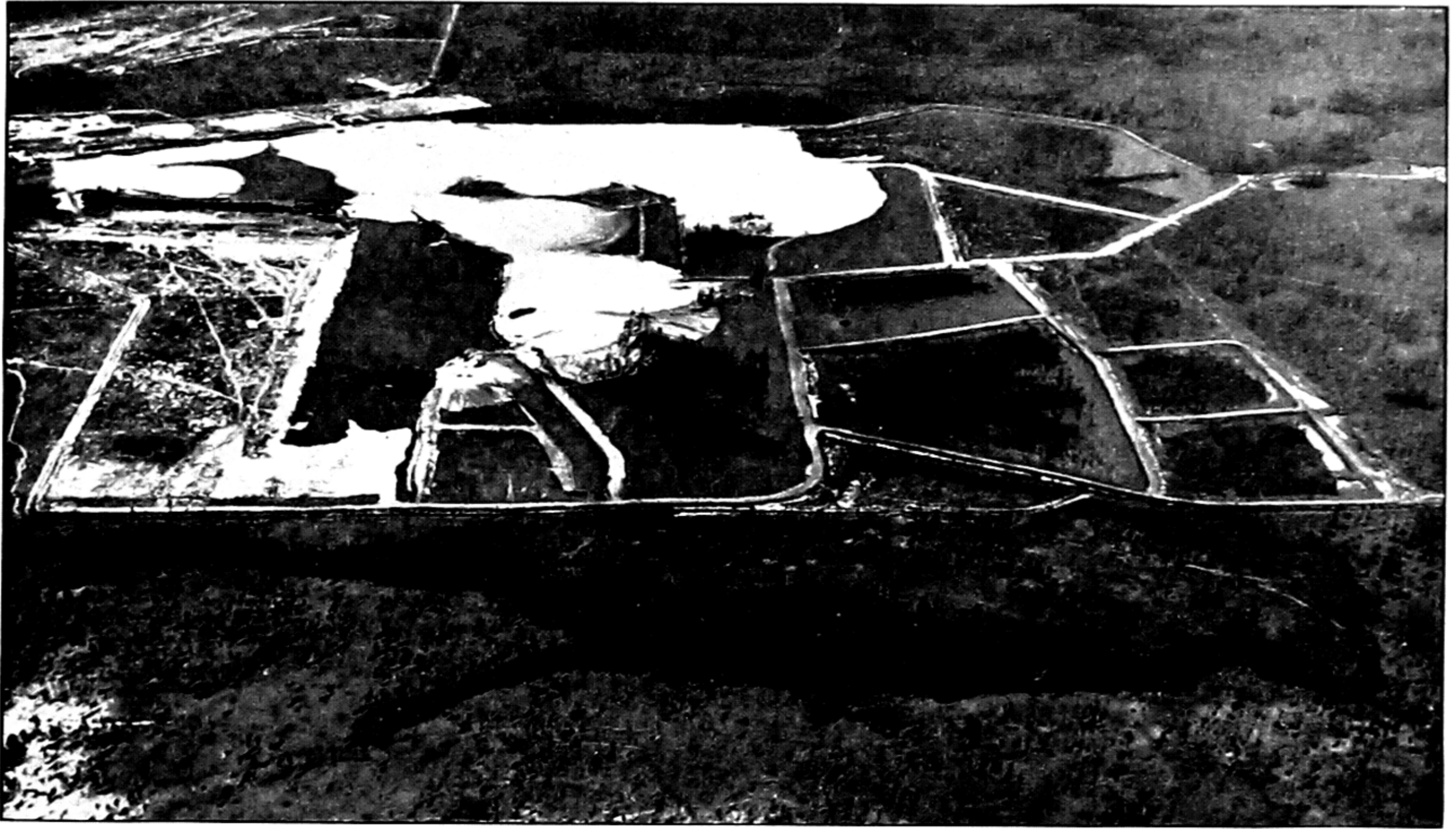
In the next 100 years many attempts were made to isolate the pure metal, but apparently without success. Zirconium is a very reactive metal and readily forms stable compounds with many elements. Efforts to reduce the mineral with carbon, boron, aluminum, iron and other agents produced only an impure form of the metal.

Impure zirconium is hard and brittle and therefore worthless as a structural material, but it does have some uses. In 1930 it came into vogue as an ingredient in a smokeless flashlight powder for photographers. Finely divided zirconium metal, mixed with magnesium and barium nitrate, provided a brilliant flash without smoke, because the zirconium ash was so heavy that it was not dispersed into the air. Instead of smudging the furniture and lace curtains with a film of magnesia powder, photographers could restrict the damage to a small deposit of zirconia ash on the rug. The popularity of smokeless powder for a time created a real market for zirconium metal. When photoflash bulbs were introduced, the bottom promptly fell out of this market. Other employments were found, however, for zirconium powder;

its relatively low ignition point, its rapid burning and its high heat of combustion make it useful in ammunition priming compounds, electric blasting caps, Very signals, airplane landing flares, movie flares and commercial fireworks.

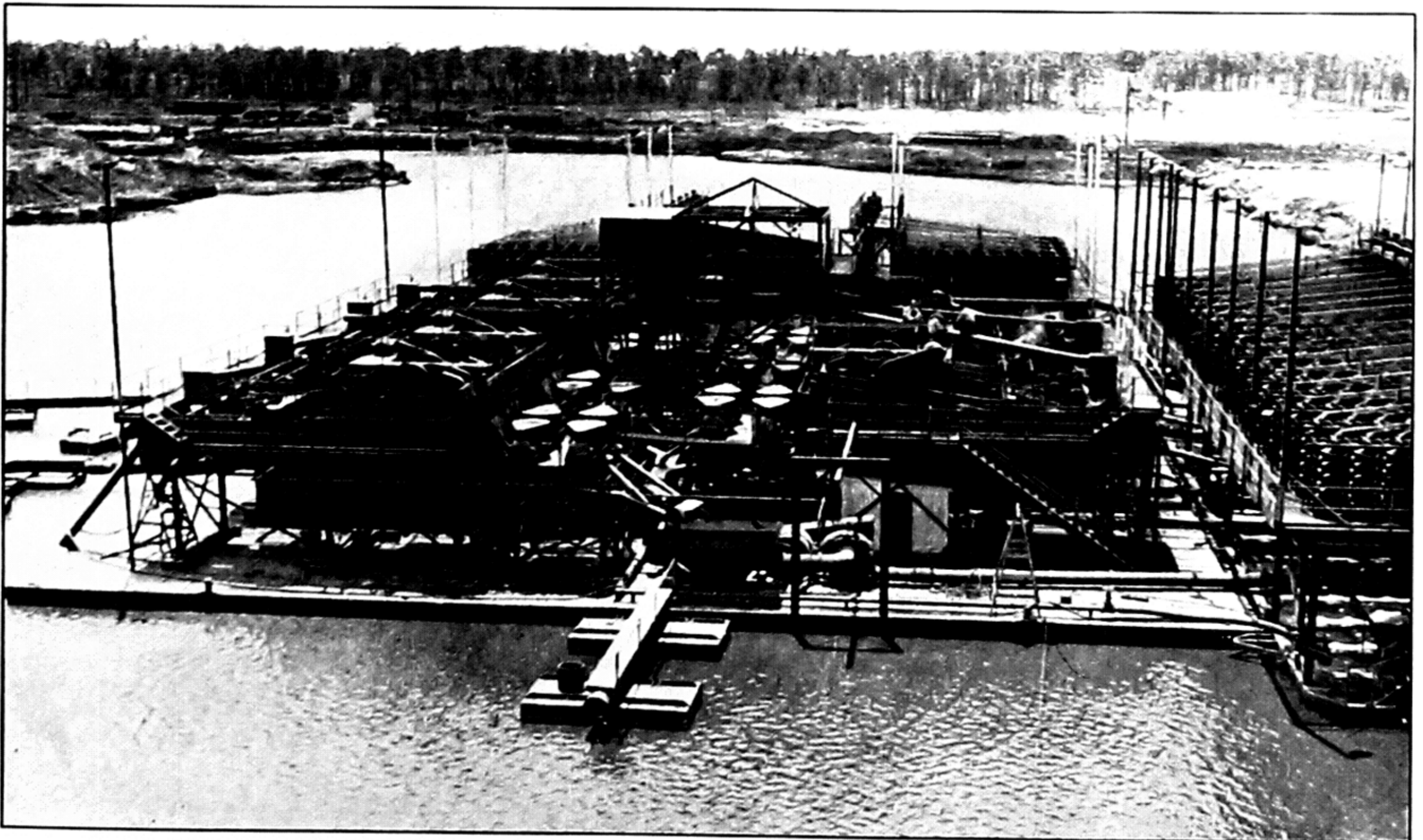
BUT it is the pure metal that has excited the great present interest in zirconium. The metal was first isolated in pure form in 1925 by the Dutch investigators A. E. Van Arkel and J. H. DeBoer, working at the University of Leyden. They succeeded by a physical method where chemical reduction methods had failed. In a Pyrex glass bulb containing a tungsten filament stretched between two tungsten electrodes they placed some crude zirconium metal and iodine. Then they evacuated the bulb with a vacuum pump and heated it in a furnace. As the bulb was heated, the iodine combined with the crude zirconium to form zirconium tetraiodide, which vaporized. This gaseous compound was then decomposed by the glowing tungsten filament, heated to a temperature of nearly 1,500 degrees Centigrade. Pure zirconium metal was deposited on the filament and the liberated iodine reacted with more of the crude zirconium in the bulb to form additional iodide. In a few hours the filament, accumulating deposits of zirconium, grew to rods several millimeters in diameter.

The metal thus purified proved to have a number of previously unsuspected properties. It was not brittle, like the impure earlier samples of zirconium, but very malleable and ductile; it could be cold-rolled, swaged or forged without requiring intermediate annealing. It resembled stainless steel in appearance, and could be polished to a mirrorlike finish. In contrast to powdered zirconium, the solid metal was extremely resistant to heat; a thin 1/16-inch sheet



ZIRCON SAND, the principal source of zirconium, is a by-product of ilmenite in this dredging operation at

Trail Ridge, Fla. Ilmenite is a mineral containing iron and titanium; it is mined here to obtain the latter metal.



ILMENITE is concentrated at the Trail Ridge plant, the tailings of which are treated for the recovery of zircon

sand. The plant is operated by the Humphreys Gold Corporation for E. I. du Pont de Nemours & Co., Inc.

can be heated to redness in an open flame without igniting, and even in an atmosphere of oxygen it will not burn below 900 degrees C. Ductile zirconium also showed remarkable resistance to attack by air, sea water, alkalis and acids (except wet chlorine and hydrofluoric acid).

Despite the spectacular properties of this wonder metal, it was slow in becoming available for wide use. For a number of years the only commercial production was by the Van Arkel process, which was taken out of the laboratory and modified for practical factory operation by the Philips Lamp Company in Eindhoven, the Netherlands. But even with improved techniques and factory-scale production, ductile zirconium metal produced by this process still costs about \$125 a pound.

In the past few years the growing realization of the metal's potentialities has spurred the metallurgical industry of the U. S. into action. At least three different processes for producing ductile zirconium, in addition to the Van Arkel method, are being developed at the present time. Of these the furthest advanced is one conceived by the renowned metallurgist William J. Kroll, who left his native Luxembourg just before the Nazi invasion in 1940 and later joined the staff at the Bureau of Mines Electrodevelopment Laboratory in Albany, Ore., where he has supervised the development of his process.

THE Kroll method, a magnesium reduction process, is carried on amidst a maze of complicated machinery and control instruments. The first step is reaction of the reddish-brown zircon sand concentrate with graphite in an electric-arc furnace. This drives off the silicon and produces zirconium carbide, which in turn is heated in another furnace and treated with chlorine gas to form zirconium tetrachloride. This product, after a treatment to remove impurities (mainly chlorides of iron), is then subjected to a reduction process which is the essential step of the method. The zirconium tetrachloride is heated in a helium atmosphere in the presence of molten magnesium metal so that the chloride reacts with the magnesium. The result is a spongy mixture of pure zirconium and magnesium chloride. By distillation under vacuum the magnesium chloride and any excess magnesium metal are removed, leaving the purified zirconium sponge. The zirconium can then be melted into metal ingots.

Zirconium is extremely difficult to handle in the molten condition, because it reduces or dissolves all of the known refractories. Consequently the melting and casting of the metal present problems. If the molten metal is not protected from air or other active gases, the cast metal produced is too hard and

brittle. The entire melting and casting operation therefore must be carried out in a high vacuum or in an atmosphere of inert gas. Two types of electric furnace have been used successfully for this operation: an arc furnace and a graphite-resistor furnace.

The Kroll process yields a very pure finished metal, containing less than .02 per cent carbon, .06 per cent iron, .07 per cent oxygen, .01 per cent nitrogen, .02 to .05 per cent aluminum and about .5 to 1.5 per cent hafnium. Incidentally, zirconium's association with the rare element hafnium is one of the intriguing aspects of its development. Hafnium, which occurs in varying amounts in all zirconium ores, is high on the list of "little-known metals" that the metallurgical industry is eager to investigate.

After preliminary laboratory experiments, Kroll and his co-workers put into operation in 1947 a pilot plant with a capacity of 60 pounds of zirconium sponge per batch. Later a larger pilot plant was built. It was so successful that in August, 1950, the Bureau of Mines completed and began to operate at Albany a full-size Kroll-process plant—the first major zirconium plant in the world. The experience in this plant indicates that zirconium ingots can be produced commercially by the Kroll method for less than \$20 a pound.

The two other developments known to be under way in the U. S. are by the Westinghouse Electric Corporation, which is trying out a process that reduces zirconium oxide with metallic calcium, and by the Titanium Alloy Manufacturing Division of the National Lead Company, which reduces zirconium tetrachloride with metallic sodium.

The cold malleability of zirconium metal is influenced to a marked degree by the amounts of oxygen, nitrogen and hydrogen that are present as impurities. Hydrogen is easy to control, because in each of the processes for producing ductile zirconium there is a vacuum step in which the hydrogen is pumped off. The presence of oxygen or nitrogen, however, is quite different. Once these gases are present there is no known method of removing them completely.

The metal is very reactive with air at elevated temperatures; this complicates initial working of the ingots. It is soft enough to be rolled, forged or swaged into almost any form if it can be protected so that it does not react with the air and become hard and brittle. The standard procedure at the Bureau's laboratory has been to sheathe the ingot in mild steel tubing with welded plugs at each end. The sheathed ingot can be heated to temperatures as high as 700 degrees C. and can be forged and rolled to a sheet less than half an inch thick. The sheath is then removed, and the worked metal, still ductile, can be cold-worked further to finished thickness.

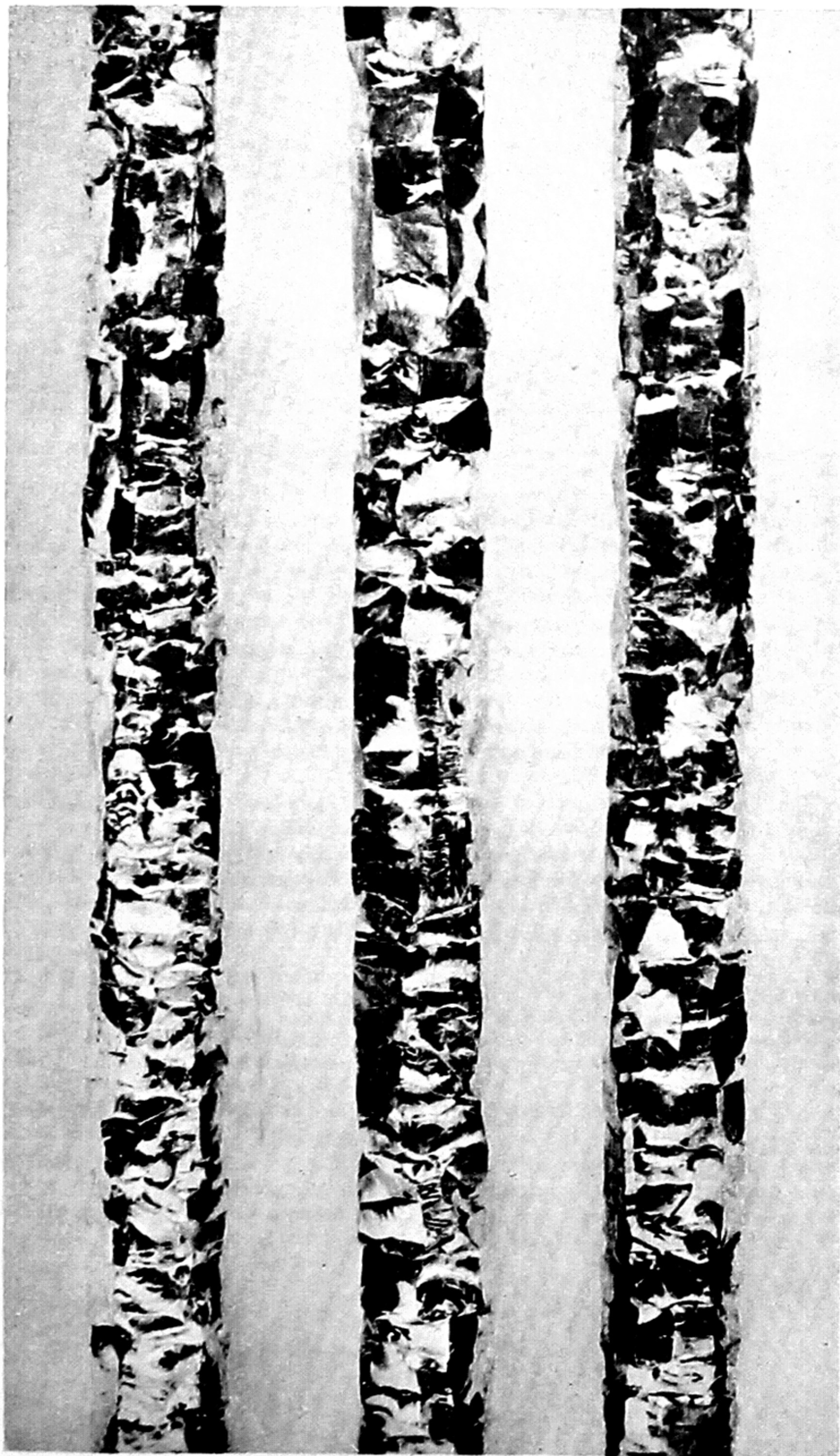
Sheet formed in this manner has an oxide layer on the surface, but this can be removed by a combination of sandblasting and a pickling operation with hydrofluoric acid and lead nitrate.

Ductile zirconium is amenable to drop-forging and pressing. Well-filled commercial shapes have been made, using dies designed for steel hand-tools. The metal can also be swaged into rod or drawn into wire.

SINCE ductile zirconium is only now beginning to become available in quantity and at something approaching reasonable cost, its possible uses are still largely unexplored. The metal is receiving considerable attention, as is already well known, from scientists and engineers who are designing reactors for harnessing atomic power. Its outstanding ability to resist attack by alkalis and acids also suggests that it will have an important future in the chemical industry, as a material for tank linings, pipes, valves, immersion heaters and the like. The metal's resistance to hydrochloric acid makes it invaluable in industries that use this acid, notably those that produce magnesium, pulp paper, salt and bleaches. Not only does zirconium equal or surpass tantalum in resistance to chemical attack, but it has the additional advantages of being lighter in weight, more plentiful and cheaper to produce.

There is a good possibility that zirconium will replace the more expensive tantalum in surgical equipment. It is well-suited for surgical instruments because of its high tensile strength and resistance to attack or discoloration by iodine, mercury salts or high-temperature sterilization. The metal has special value for suture wire, bone screws and cranial plates. Tissue grows directly to the metal without any signs of decay or deterioration; body fluids will not corrode it. One medical use to which zirconium has already been put is in the construction of a false eyeball. The rear half of the eyeball is made of zirconium, with hooks that attach to eye muscles. This device can be moved by the muscles and is indistinguishable from a normal eye. At the University of Pennsylvania investigators have been experimenting with zirconium to find out whether it can solve a problem created by the removal of a diseased lung. In such cases other organs frequently move into the space left by the removed lung and cause trouble. The experimenters formed finely rolled zirconium sheets into artificial lung spacers and spliced them into laboratory dogs. After three years the dogs are still frisking around in perfect health.

Zirconium has recently been put to use as the light source in a new lamp which has a brilliance nearly rivaling that of the sun. This high-power, high-intensity



ZIRCONIUM RODS were made by the Van Arkel process, in which crude zirconium is allowed to react with iodine to form zirconium tetraiodide. This gaseous compound is then decomposed by heating to form pure zirconium.

electric arc lamp heats a film of molten zirconium to a temperature near 6,500 degrees Fahrenheit. The light source is only a fifth of an inch in diameter. It operates in the open air, not in a glass bulb. In a 1,000-watt lamp, operating at 55 volts and 18 amperes alternating current, the light has 20 times the brightness of the ordinary tungsten filament and one-eighth the brightness of direct sunlight. Possible applications for the lamp are in motion-picture projection, television, lithography and photocopying.

ZIRCONIUM'S affinity for gases provides another possible use for it in radio tubes. These tubes have a small piece of metal enclosed in them to help produce a higher vacuum. After a tube has been evacuated to a certain pressure by a vacuum pump, the metal, called a "getter," is volatilized by heat and in the gaseous state combines with the residual gases in the tube and takes them out of circulation. As a getter zirconium may be able to act over the entire life of the tube, absorbing any gases that may be liberated from the walls of the tube or that may leak into the tube. Thus it would increase the life and efficiency of the tube. Zirconium probably will be used similarly in fluorescent lamps.

If the chemical activity of the metal can be controlled, it will have wide application as a cleanser in metallurgical processes. It would be an ideal desulfurizing and deoxidizing agent for steel and at the same time would tie up the nitrogen and carbon, so that the steel would not age. The particles formed in a steel by zirconium—mainly nitrides, oxides and sulphides—produce nuclei; therefore the metal would refine the grain of the steel.

As an alloy, zirconium promises to be invaluable. Zirconium boride can withstand temperatures up to 6,000 degrees F., so it is a good candidate for use in rocket combustion chambers. Alloys of zirconium and copper harden with age. Zirconium imparts desirable properties to magnesium; when used in high proportions it makes the magnesium structurally strong and in lower proportions refines the grain size and improves the workability. A gold-zirconium alloy is so hard and resistant to corrosion that it may replace platinum and iridium in fountain-pen tips. These two metals also combine to form what is probably the hardest gold-base contact alloy known.

Potentially zirconium's uses are legion. Like all the new metals of the past, it will no doubt be adapted eventually to uses not dreamed of today.

The Author

STEPHEN M. SHELTON is Regional Director of the U. S. Bureau of Mines in Albany, Oregon. Born in 1903 in Bennettsville, South Carolina, Shelton went to Yale and received his M.A. from George Washington University in 1931.

Bibliography

ZIRCONIUM METAL, AS OF 1949. R. I. Jaffee in *Journal of Metals*, Vol. 1, No. 7,

pages 6-9; July, 1949.

PRODUCTION OF MALLEABLE ZIRCONIUM ON A PILOT-PLANT SCALE. W. J. Kroll, W. W. Stephens and H. P. Holmes in *Journal of Metals*, Vol. 188, No. 12, pages 1,445-1,453; December 1950.

MELTING AND CASTING ZIRCONIUM METAL. W. J. Kroll and H. L. Gilbert in *Journal of the Electrochemical Society*, Vol. 96, No. 3, pages 158-169; September, 1949.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

THE GROWTH OF CRYSTALS

by Robert L. Fullman

How do atoms or molecules in a vapor or solution form the regular architecture of a solid? Much can be explained by the assumption that they fall into place rather like the bricks laid in a spiral ramp.

All metals, indeed nearly all the basic solid materials of modern technology, are made of crystals. Hence an understanding of how crystals are built—how they grow—is an important aspect of the science of materials. Like many other seemingly simple phenomena, the construction of a crystal is not as neat and orderly as it looks at first sight. Twenty-five years ago physicists thought they had a reasonable notion of how crystals grow. Then an elementary experiment disabused them. It led to a series of surprising discoveries and to a new picture of crystals.

Nature builds a crystal somewhat as a bricklayer builds a brick wall. From a jumble of atoms or molecules heaped in no particular order (*i.e.*, in gas or liquid form) it takes suitable bricks one by one and stacks them in a regular array. If we picture the units (atoms or molecules) as little cubes and suppose that they line up consecutively and stay put, a perfect crystal in course of construction would look like the illustration at the left below. But of course atoms or molecules ordinarily have kinetic energy and are in random thermal motion; they may light anywhere on the crystal sur-

face, and quickly take off again. Thus even in this idealized picture the growth of the crystal is not so orderly as the laying of bricks but proceeds in a more haphazard fashion, illustrated by the second drawing [*at the right below*]. A unit that alights on the crystal surface is likely to stay bound to the crystal only if it migrates to a niche where at least three of its six faces are bonded to units already fixed in the crystal. Moreover, a layer needs a certain minimum number of units to assure growth, for if too large a proportion of its members are at corners or edges, they will not stick.

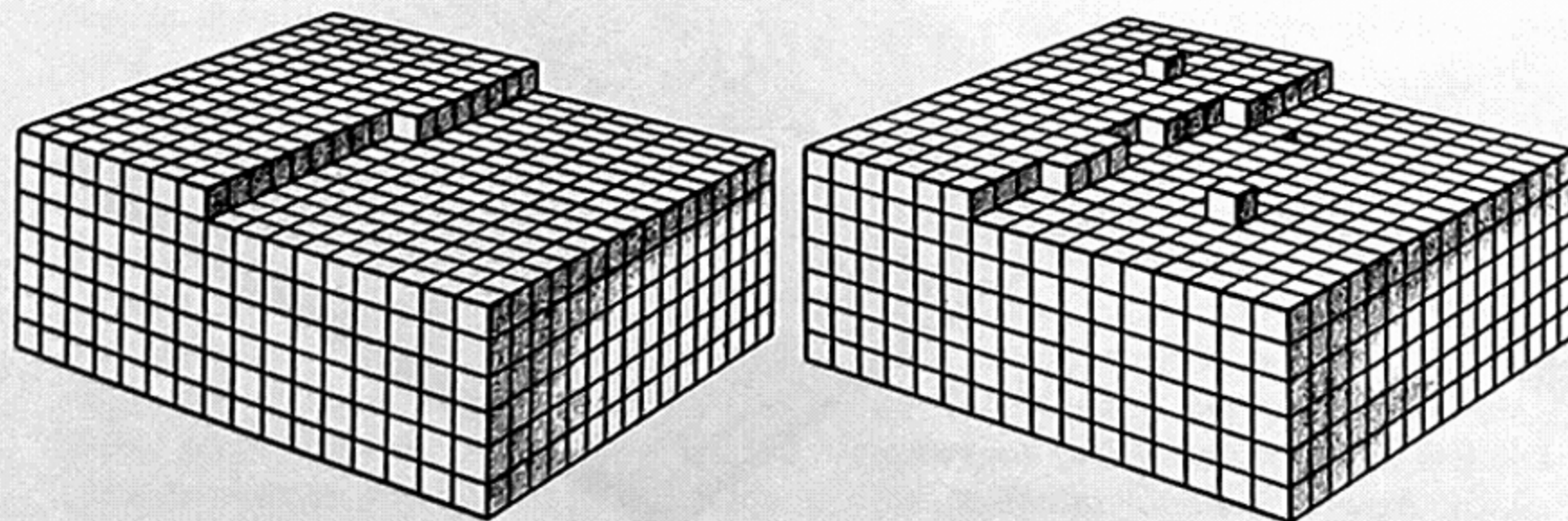
The rate of escape from a solid to a surrounding vapor depends on the intensity of the molecules' motion—*i.e.*, on the temperature. The rate of return from the vapor to the solid depends on the concentration of molecules in the vapor—*i.e.*, on the vapor pressure. For every temperature there is a "saturation vapor pressure" at which the rates of escape and of deposit at a step balance. Under these conditions the crystal does not grow. It can grow only when the vapor is supersaturated. This model suggests, further, that after a layer on the crystal surface has been filled, continued growth

becomes more difficult; that is, it should take a higher degree of supersaturation to start a new layer than to fill one already well started.

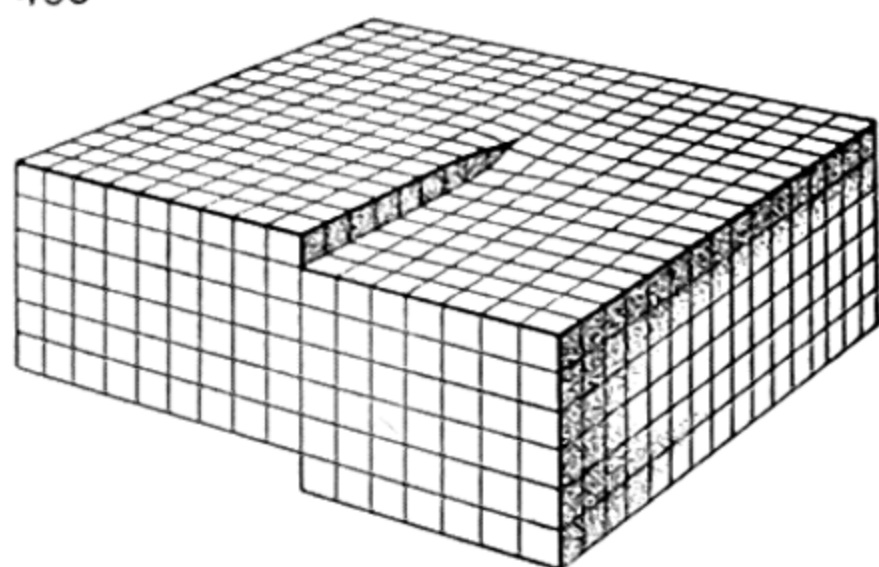
It was at this point that the classical model of crystal growth fell apart. Calculations had indicated that the formation of islands, or nuclei, large enough to start new layers, so that a crystal could grow, required a vapor pressure at least 25 to 50 per cent greater than saturation. But in 1931 the German physicists M. Volmer and W. Schultze succeeded in growing crystals from vapor with a supersaturation of only 1 per cent. The theoretical prediction of the rate at which a crystal should grow at this saturation turned out to be too small by a factor of 10^{1000} —which may well be the all-time record for discrepancy between theory and experiment!

How could a crystal initiate and build new layers so easily? Physicists went back over their analyses and calculations, but they could find no answer that made sense on the basis of the old model. The problem remained baffling until, in 1949, an English physicist suggested a new model. F. C. Frank, of the University of Bristol, pointed out that the growth of a crystal could easily be accounted for if one supposed that its layers were not laid independently one upon the other but were built spiral fashion by a continuous process. This could occur if the crystal contained an imperfection of the type known as a "screw dislocation."

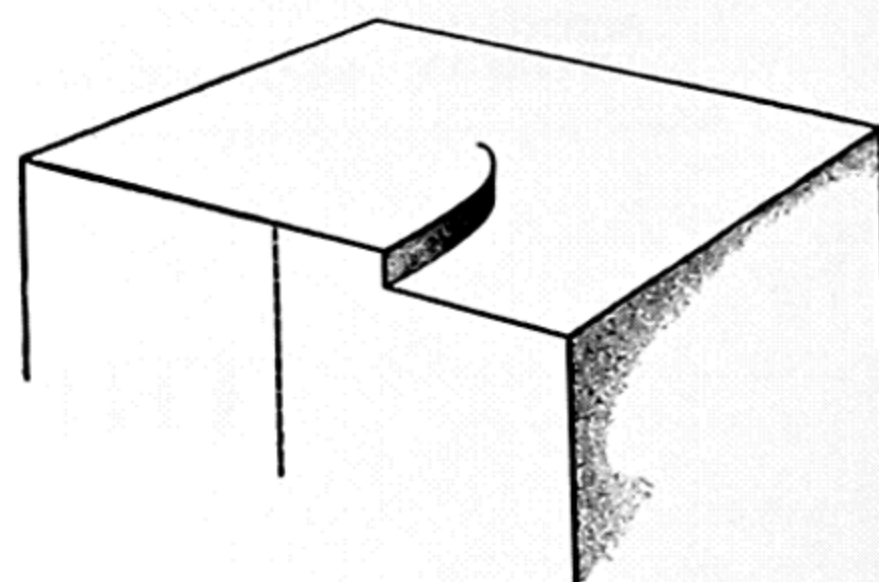
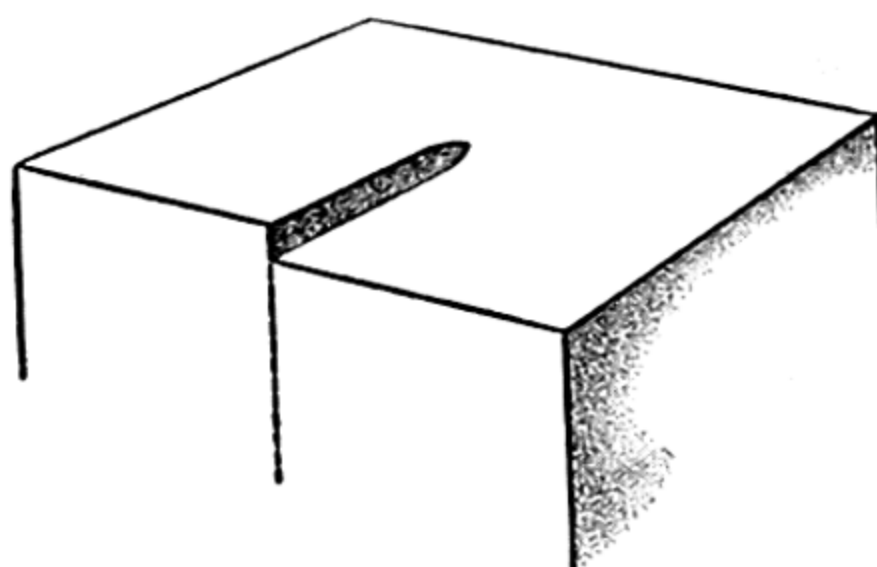
Imagine a crystal lattice in which two adjacent sections are sheared along part of their length, so that one partial section is displaced a full layer below the other [*see first diagram on page 496*]. The reason this is called a screw dislocation can be illustrated by the fact that if a tiny observer were to walk in a circle



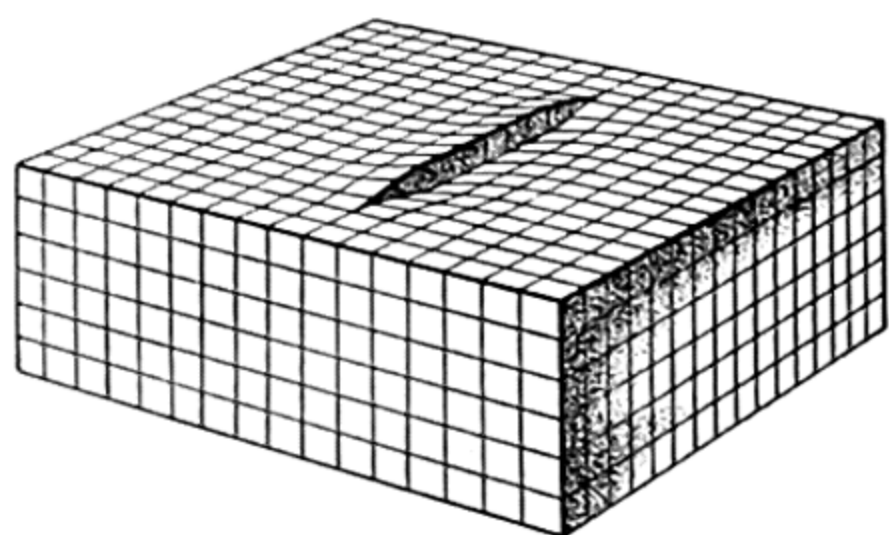
GROWING PERFECT CRYSTAL at an arbitrary point in its development would appear as at left, if its molecules had no thermal motion. Drawing at right shows the effect of this motion. Each cube in these highly schematic drawings represents a single molecule.



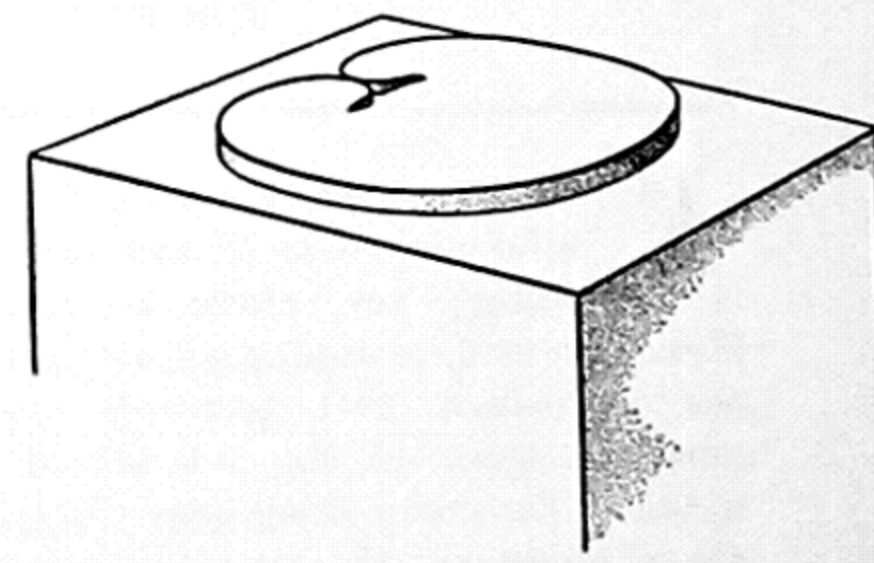
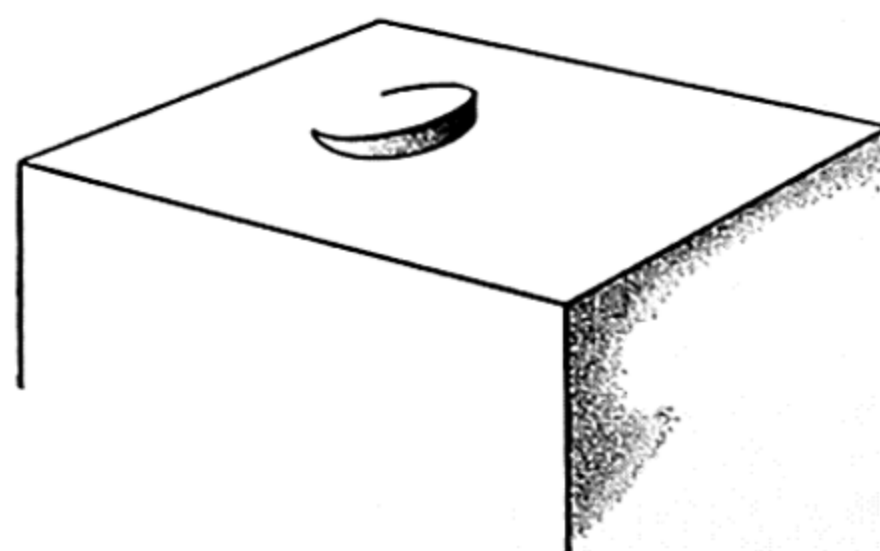
SCREW DISLOCATION, illustrated at left, produces its spiral growth steps in the manner indicated by the succeeding diagrams. A screw dislocation is the type of imperfection that would result if a cut



were made part way through a crystal and the two sides slipped over one another. The result is a permanent step extending across a portion of the crystal face and anchored at the boundary of the cut,



CLOSED GROWTH LOOPS can arise from a step which begins and ends within the surface (*left*) connecting a right-handed screw dislocation with a left-handed one. This step is anchored at both ends and can



grow only by bulging. The second drawing shows the bulging step just beginning to turn back on itself. In the next figure its parts have met behind its original position. Finally the inner section detaches from

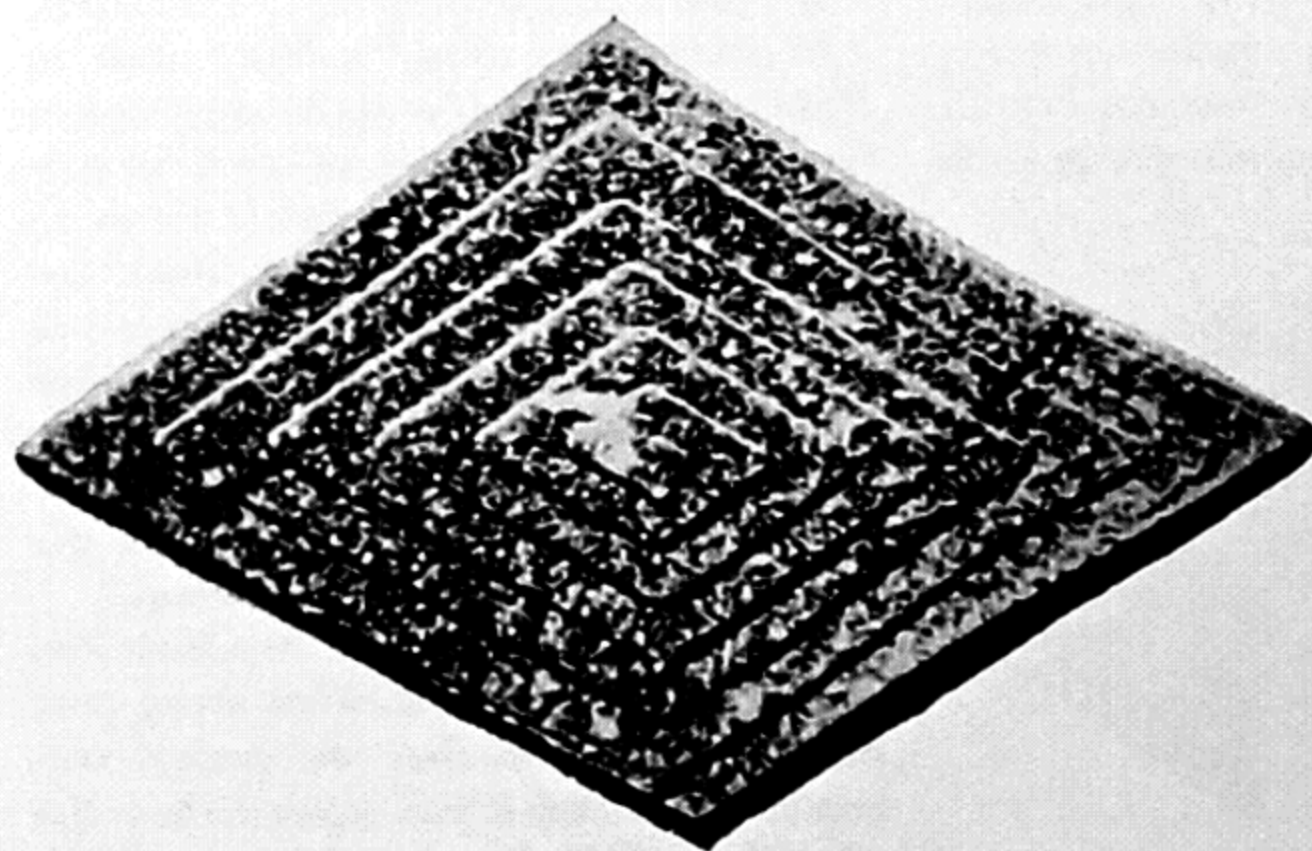
around the end of the cut, starting at the top, he would find himself one level lower at the end of each circuit, as if he were descending a spiral (actually helical) ramp. The model shown here is known as a right-handed screw dislocation because the path spirals downward to the right.

Now Frank realized that such a dislocation would provide a constantly open step at which atoms or molecules could easily gain a foothold and build the crystal without ever having to start a

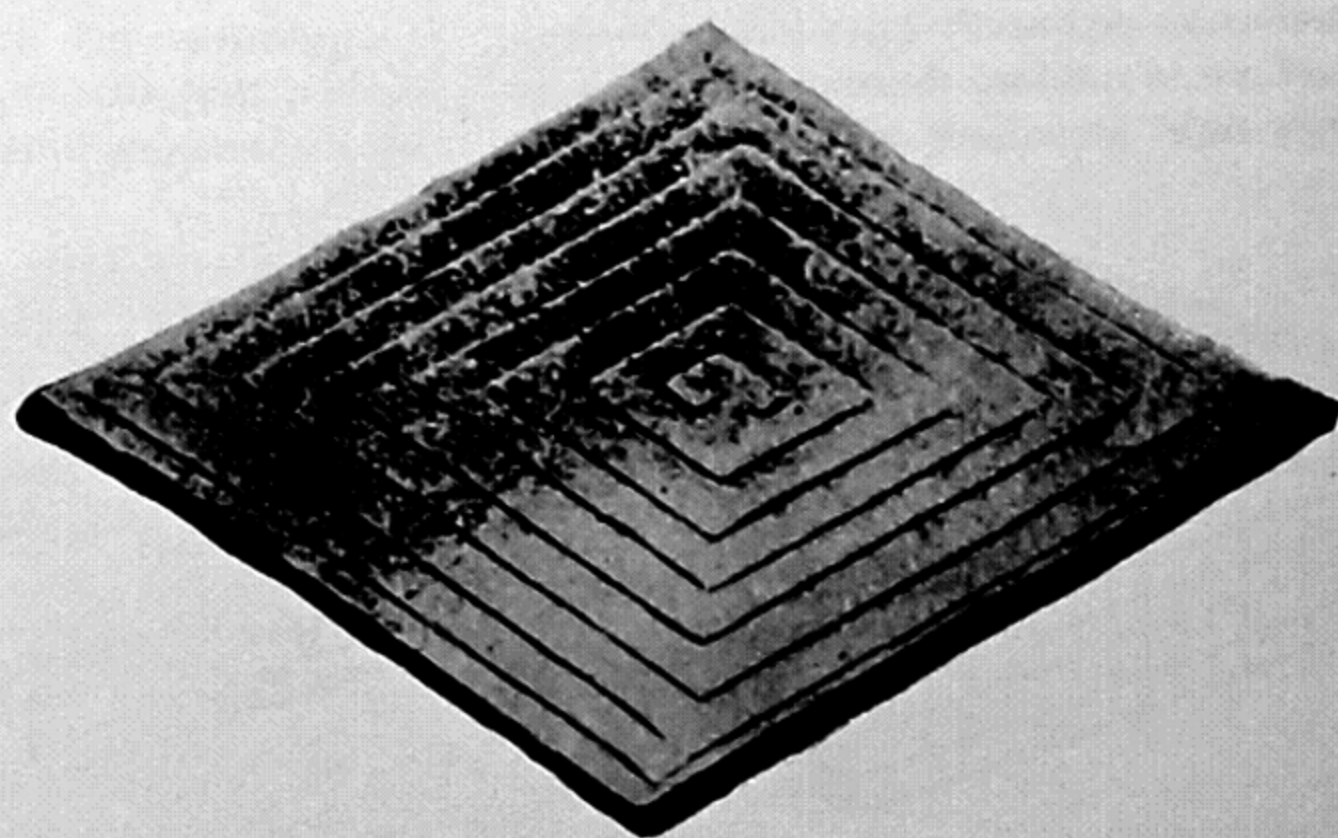
new layer. The step would sweep around the crystal face as new units were added to it. But in the process it would bend [see drawings in upper row above]. To understand why this is so, consider a line of skaters pivoting around one end. In order to keep a straight line, the skaters at the outer end must go faster than those nearer the pivot, because they have farther to travel. In the case of the crystal step, the outer end does not build faster, for units attach themselves at random, which is to say at a more or

less uniform rate, along the length of the step. Thus in the build-up the swing of the outer part of the step around the circle falls steadily behind, and the step bends to form a spiral.

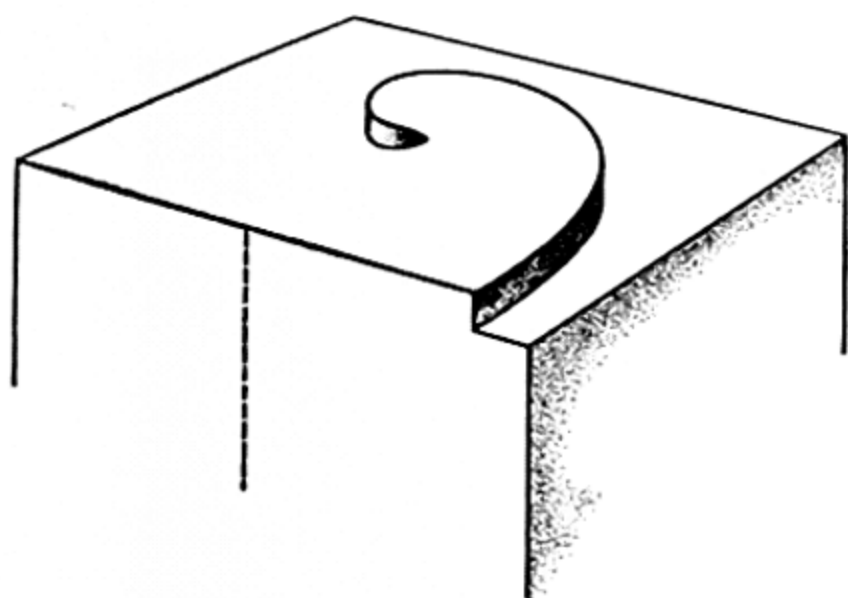
Was Frank's radically new picture of crystal growth correct? It did not have to wait long for confirmation. At the meeting of scientists where his theory was presented, L. J. Griffin of the University of London showed some photomicrographs of crystal surfaces exhibiting the predicted spiral step patterns.



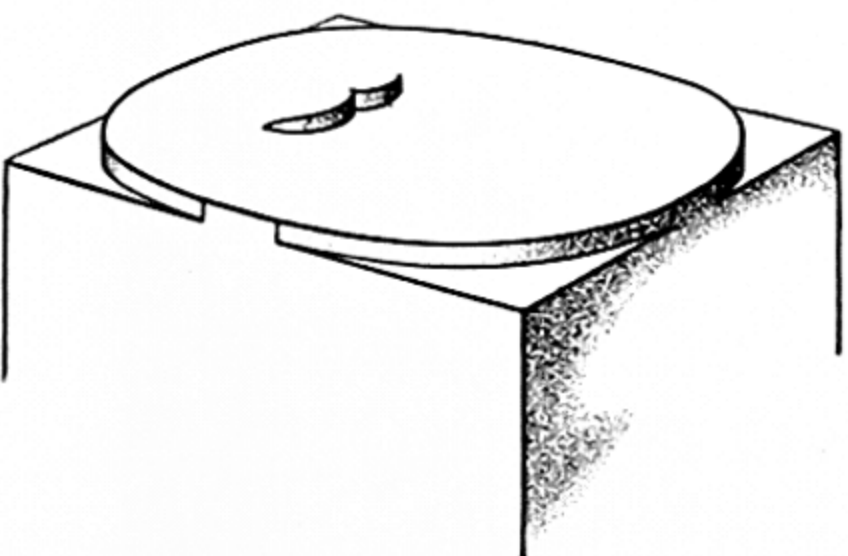
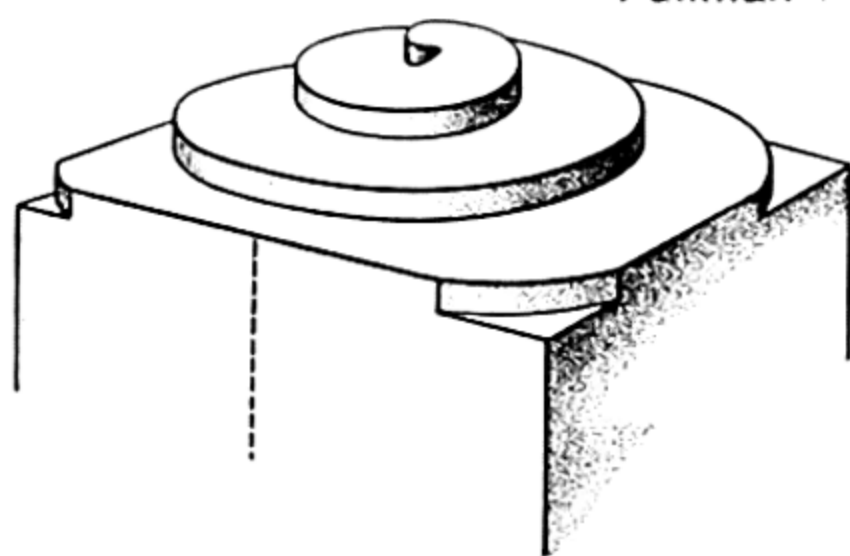
GROWTH STEPS on paraffin crystals are revealed by photomicrographs. Spiral (*left*) is broken up into a series of straight segments, and the loops (*right*) are polygons. The reason for this de-



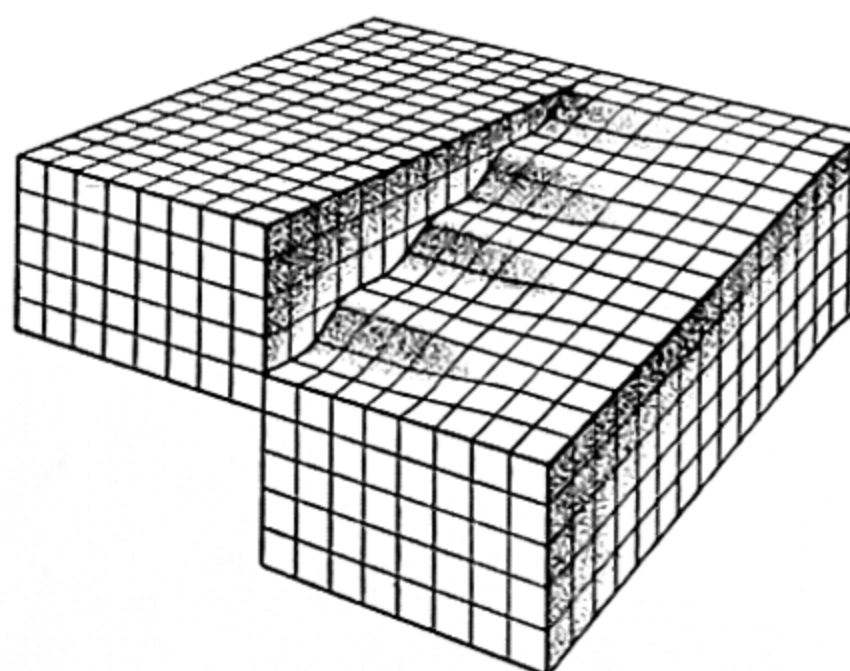
parture from circular form predicted by the theory is that these crystals do not grow equally fast in all directions. The straight sections advance in the directions of slowest growth along the surface.



Molecules depositing on the surface from a vapor would lodge against the step, causing it to advance. Since one end is fixed, the step would pivot around that point, with the outer points falling behind the inner, producing a never-ending spiral layer on the surface.



the completed layer and shrinks inward until it becomes a new straight step located on top of the growth loop first formed.



FOUR-LAYERED STEP is formed by four closely-spaced dislocations of the same kind. Larger groups give many-layered steps.

What is more, since then intensive examination of crystals under the microscope has brought to light various growth patterns which can only be explained, indeed were predicted, by the screw dislocation theory.

Nature rarely makes a perfectly regular crystal. There is usually at least one dislocation—as we have seen, it is a necessary condition for growth at low supersaturation. Generally a crystal has more than one dislocation. The presence of several dislocations complicates the crystal's growth pattern. For instance, two neighboring dislocations, one right-handed and one left-handed, may form a step anchored at both ends [*first drawing in second row on opposite page*]. As new units build onto this step, it takes a semicircular shape [*second diagram*]. Eventually the build-up produces a ring-shaped step [*third diagram*] which then builds out to complete a layer on the crystal. But in the meantime a new step has been formed on top of the original one, and it becomes the nucleus for a new layer [*fourth diagram*]. As the process repeats, the crystal grows in layers which are closed rings rather than spirals. A photograph of such a crystal is shown on the opposite page [*bottom right*]. The steps here are straight-sided

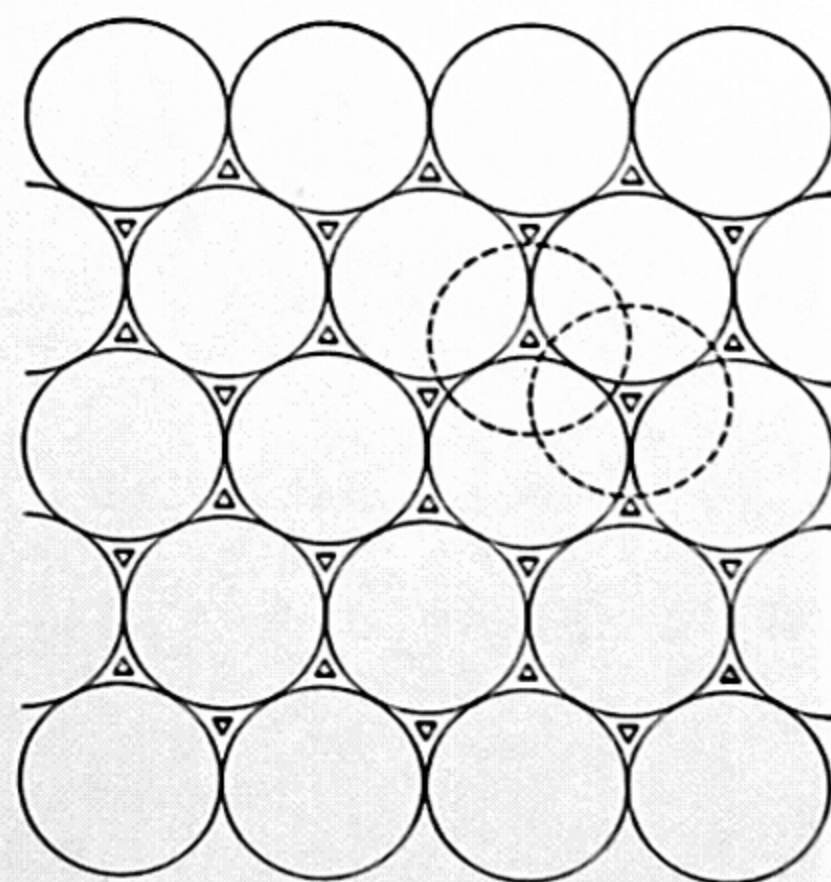
polygons, rather than curved rings, because growth has not proceeded evenly all around the periphery.

The height of a crystal step can be measured. Usually the step is just one unit high. But some crystals have been found to have steps many layers high. The theory suggests that multilayered steps may arise from a group of dislocations which are close together and all of the same sense—right-handed or left-handed. Each dislocation contributes one layer to the thickness of the step. The photographs on the next page show a cadmium iodide crystal with a step many layers high. The short, straight line at the center of the spiral marks the initiating array of screw dislocations.

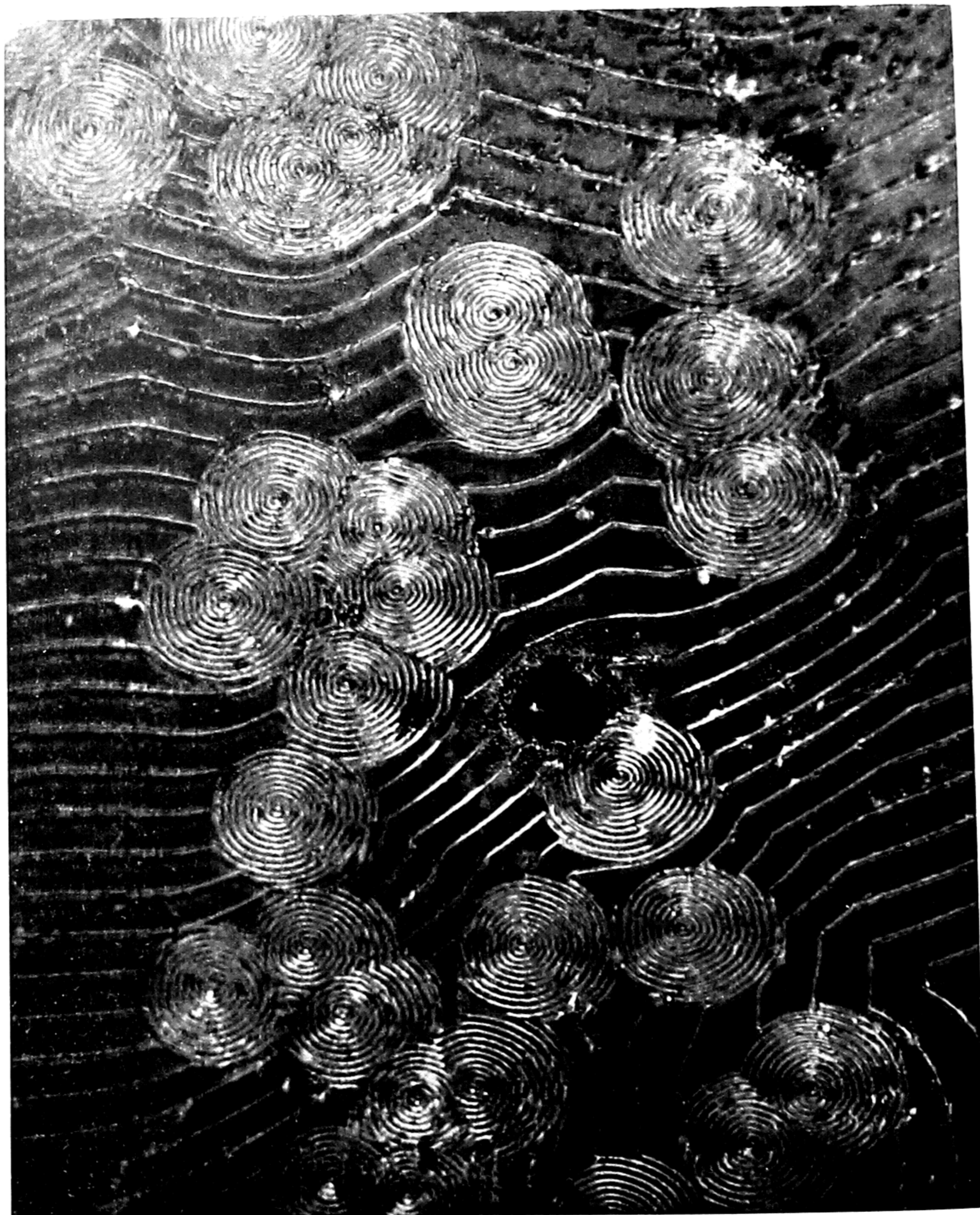
Here the theory helps to clear up a peculiarity of crystal structure which has long mystified crystallographers. It has to do with the stacking of layers in a crystal. To follow the reasoning we must discard the oversimplified picture of building units that we have been using for illustration so far. The building blocks of a crystal—its atomic or molecular cells—are not cubes but more like spheres. Certain crystals have a close-packed structure, which means that the spherical units are piled not one directly above the other but in the manner of a mound of cannon balls, with

each ball nestling in a pocket between its neighbors. Now a look at such an array shows that each successive layer may be stacked on the one below in one of two positions. In the diagram at the right, viewing a layer from above, the positions are marked Δ and ∇ . When a layer is placed on this one, close packing requires that all the balls be laid either in the positions marked Δ or in those marked ∇ . The next layer again may be stacked in either one or the other set of positions. If the layers are stacked alternately in one position and then the other, the sequence is denoted by $\Delta\nabla$. But crystals often stack up in much more complex patterns. For example, one kind of silicon carbide crystal has the pattern $\Delta\Delta\nabla\nabla\nabla$, $\Delta\Delta\Delta\Delta\Delta$, etc. Here the group forming the repeat in the pattern consists of five layers. In some crystals of silicon carbide the repeat unit has been found to be as large as 29 layers.

How can information about the pattern be transmitted over so great a distance; that is to say, how can each unit be matched to one 29 units away? Classical crystallography could suggest no explanation other than the possibility of some unknown long-range force between molecules. The dislocation theory of crystal growth now provides a simple, though incomplete, answer. A series of screw dislocations may generate a step many layers high, which then becomes the pattern unit that repeats again and again as the crystal grows. One aspect that the theory does not explain is the fact that only 14 such variations of the crystal pattern have been found in silicon carbide; the theory suggests that there should be an almost unlimited

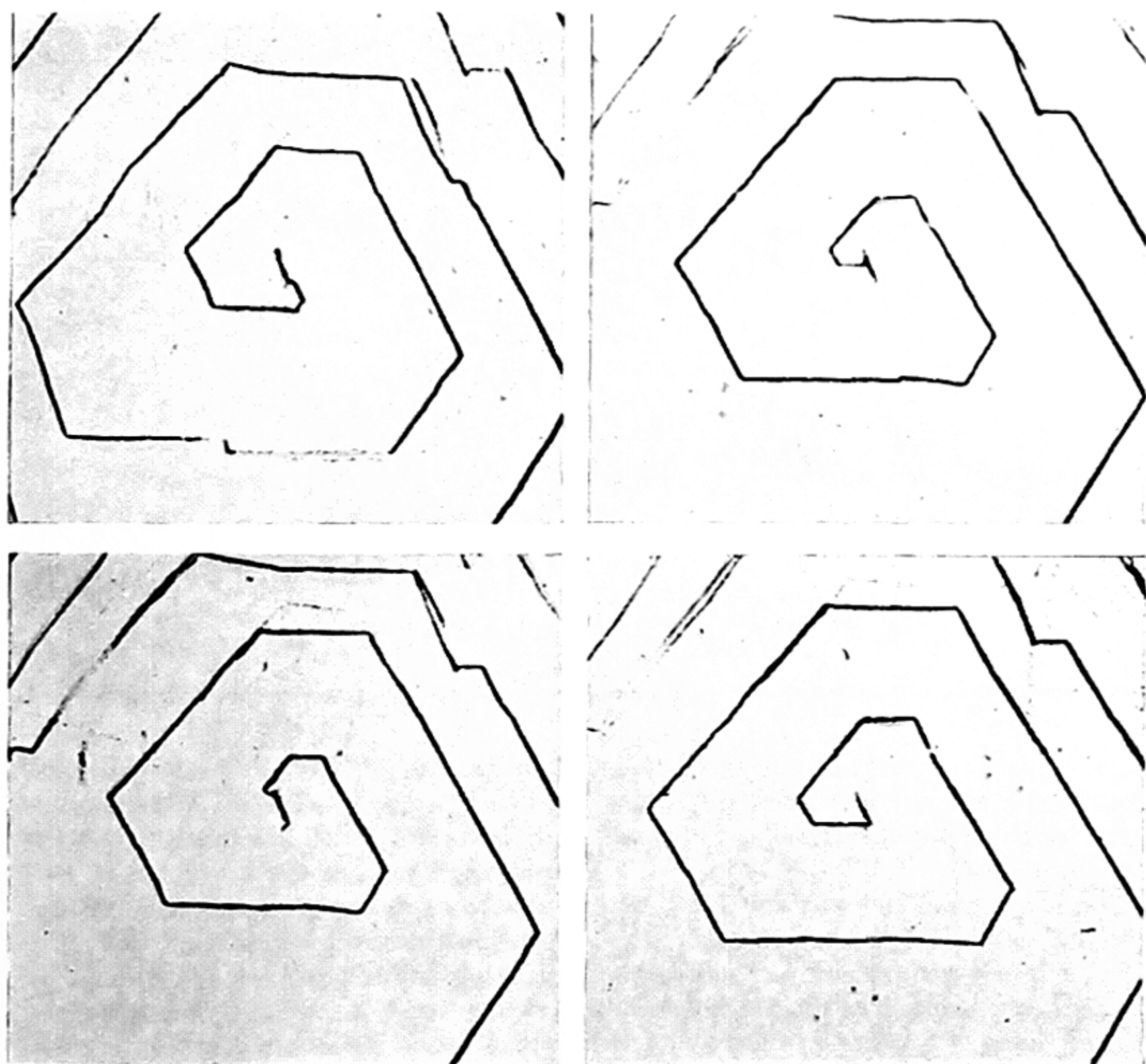


CLOSE-PACKED molecules are represented by circles, with possible positions for members of the next layer shown dotted.

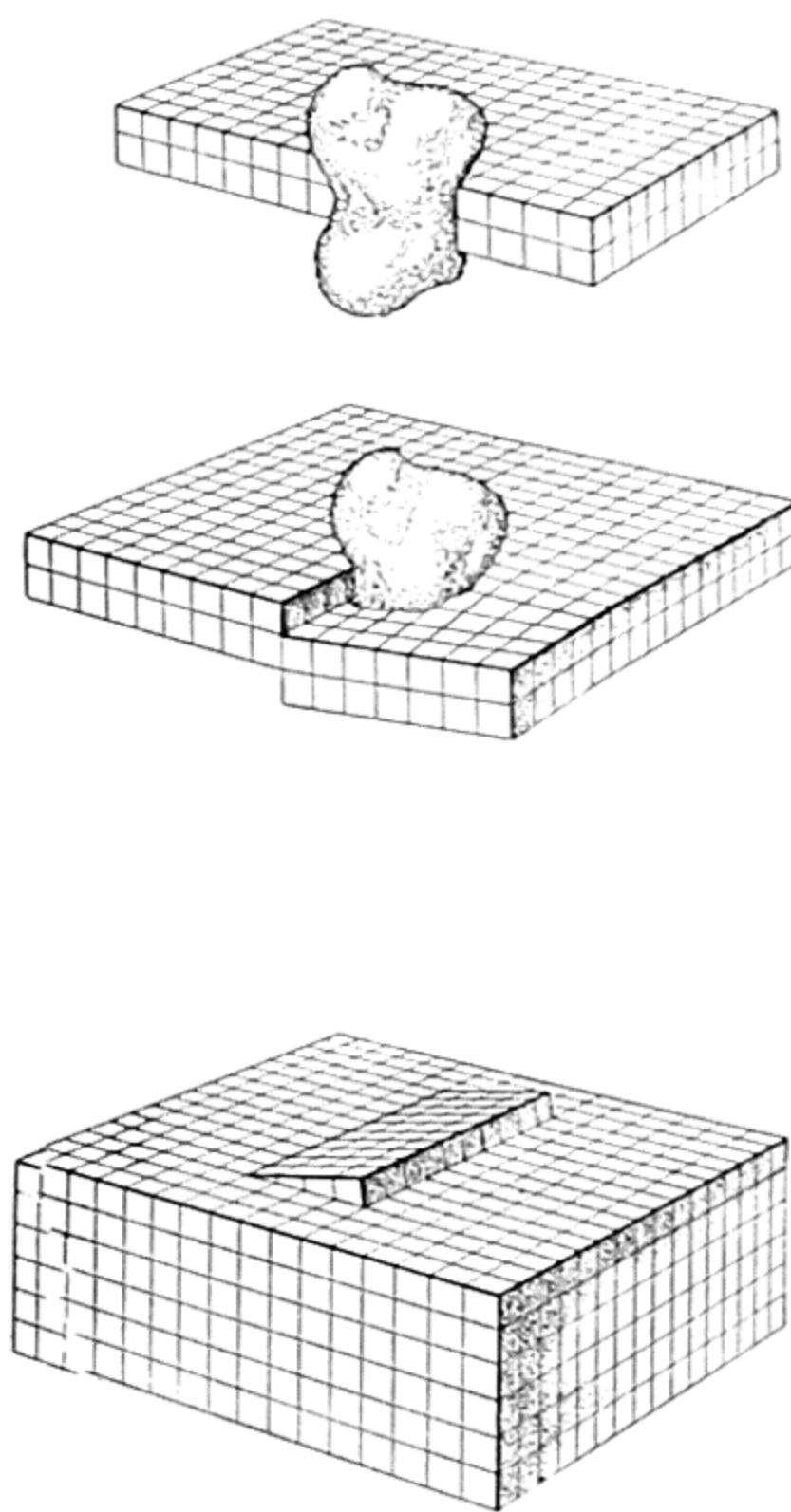


CRYSTAL GROWTH SPIRALS on the surface of a silicon carbide crystal are seen magnified 250 diameters. Each spiral step origi-

nates in a defect called a screw dislocation. From the side, the surface looks like an ascending ramp wound around a flat cone.



GROWING SPIRAL is seen in four successive positions on a crystal of cadmium iodide. This step, which is many layers high, is produced by a large group of screw dislocations.



LOOP SOURCE could be formed if growing crystal picked up a foreign particle tilted with respect to its molecular layers.

number of possible patterns in the crystal structure.

How do screw dislocations arise in the first place? This is still a largely speculative question. Almost all crystals begin their growth on some small foreign nucleus, just as raindrops condense on dust particles. Some of these nuclei may have surface irregularities which correspond to screw dislocations and are passed on to the growing crystal. Other

crystals may pick up dislocations as they grow. A strong mechanical disturbance may cause a thin growing crystal to buckle and shear; indeed, thin crystals have been induced to grow by pressing a fine needle point against them. Dirt particles picked up during growth also may introduce dislocations.

The preparation of metals and other crystal materials in useful sizes without weakening dislocations is a still unsolved

problem. If ways can be found to circumvent the restrictions nature normally places on the growth of crystals, truly revolutionary materials will become possible. Whether or not this goal is achieved, the dislocation theory has shed light on a number of properties of crystals—their strength, the rates at which atoms diffuse through them and the rates at which they grow.

The Author

ROBERT L. FULLMAN is a research associate at the General Electric Research Laboratory, where he has been since 1948. Born in 1922 in Sewickley, Pa., a suburb of Pittsburgh, he went to Yale University to study metallurgy and became interested in the structure of crystals. He was the codiscoverer, with Arno Gatti, of the method for producing perfect iron crystals. During World War II Fullman served in the U. S. Navy for three years and did duty on a destroyer in the Pacific.

Bibliography

- CRYSTAL GROWTH AND DISLOCATIONS. F. C. Frank in *Advances in Physics*, Vol. 1, No. 1, pages 91-110; January, 1952.
- CRYSTAL GROWTH AND DISLOCATIONS. Ajit Ram Verma. Academic Press, Inc., 1953.
- DISLOCATIONS IN CRYSTALS. W. T. Read, Jr. McGraw-Hill Book Company, Inc., 1953.

A MODEL OF THE NUCLEUS

by Victor F. Weisskopf and E. P. Rosenbaum

As an aid to understanding the atomic nucleus, physicists visualize it in terms of simplified models. A surprisingly fruitful approach is to regard it as a cloudy crystal ball.

Modern physics is frequently accused of deserting the real world for abstract mathematics. Instead of attempting to explain nature in terms of what we can see and feel, the impeachment runs, theoretical physicists offer only an arid set of equations whose physical meaning they will not even think about, let alone interpret to the vulgar.

The reproach is not deserved. Nearly every physical theory begins with some model by which the physicist seeks to interpret it. It is true that as a theory grows more sophisticated the model tends to lose its connection with everyday experience. But the physicist never

ceases to be conscious of the fact that his equations represent an endeavor to account for the behavior of a real, physical universe.

The field of nuclear physics is still an unfamiliar territory; hence model-building is a major occupation of its investigators. In the nucleus of the atom we are dealing with an entirely new kind of matter: an unimaginably dense material composed of protons and neutrons, collectively called nucleons. How these parts interact, what law governs the enormous force they exert on each other—these are still mysteries. Therefore we cannot yet make use of the elegant mathematical methods of quantum me-

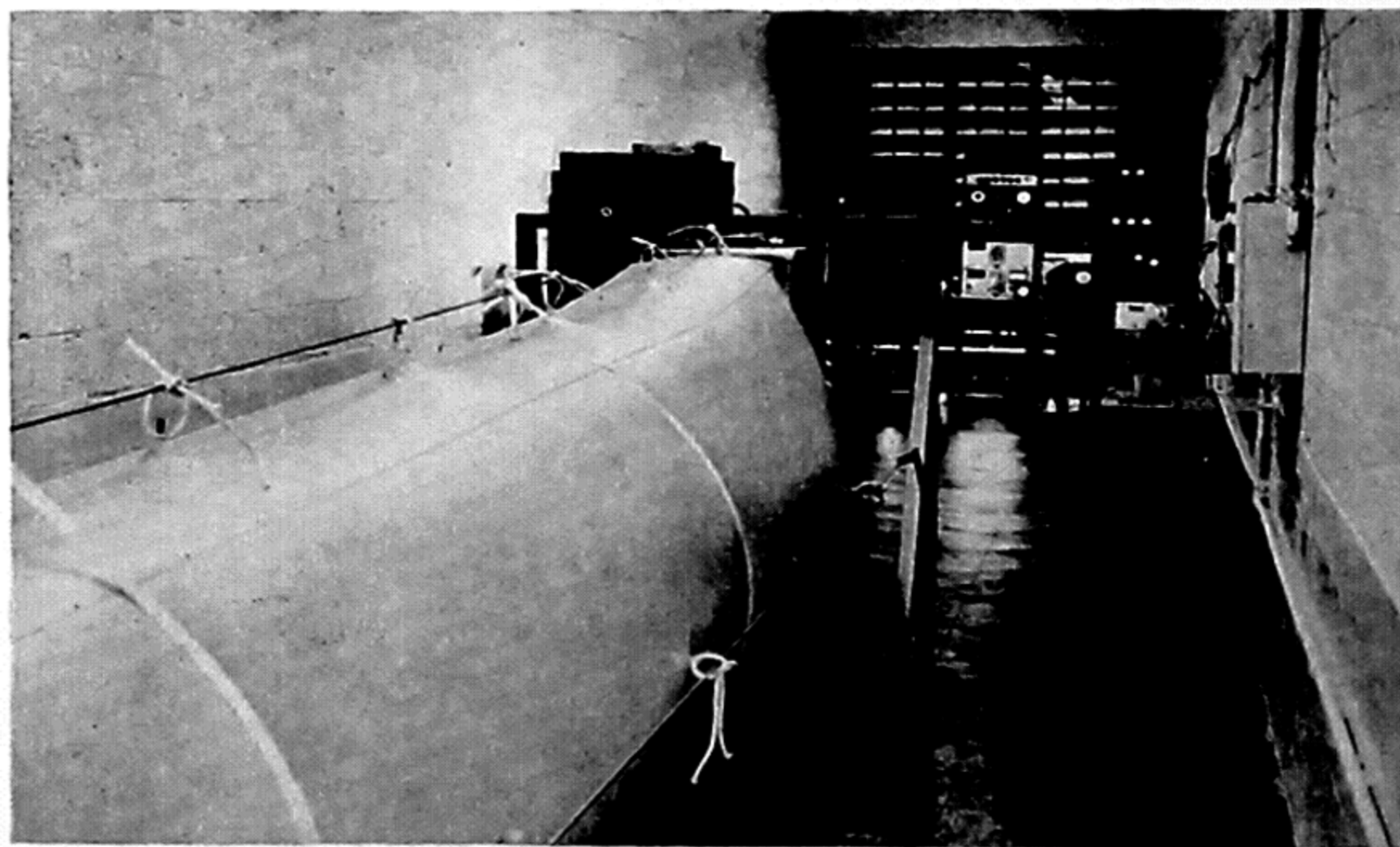
chanics, which have so successfully explained the properties of ordinary matter, involving only electrons. Until the force acting inside the nucleus is as well understood as the electrostatic force that acts on the atomic electrons outside, the equations for nuclear matter cannot be written. And so, to help organize the facts learned by experiments in nuclear physics laboratories, and to guide the imagination toward useful insights, nuclear physicists have recourse to various simple, intuitive models picturing a hypothetical structure of the nucleus.

Every model we build is an effort to make sense of some particular class of experimental results. In our attempts to get at the invisible world of the nucleus we are in the position of the three blind men examining an elephant: different experiments bring out different aspects of the subject. Hence our different models are not mutually exclusive. As we shall see, they can often be combined to give a more complete understanding.

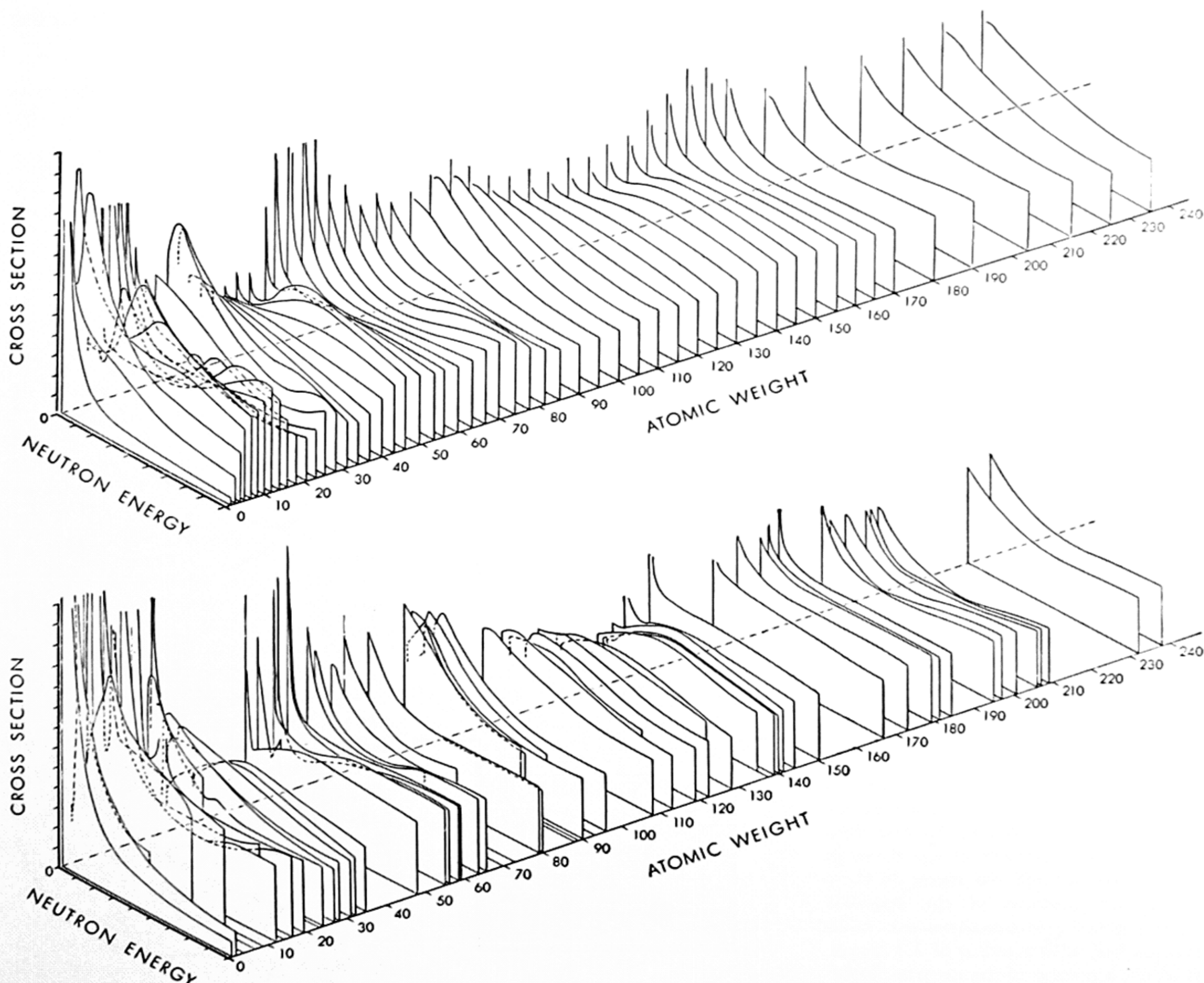
We shall describe here a very simple model of the nucleus which has been surprisingly successful in predicting some quite complex behavior. It sidesteps the question as to how the individual nucleons may interact, concentrating instead on what might be called the gross structure of nuclear matter. To understand how the model works we should first know something about the specific experiments it was designed to illuminate.

The experiments have to do with bombardment of the nucleus with neutrons. These uncharged particles can easily enter the nucleus, because they are not resisted by the electrostatic force of the nucleus' positive charge.

Consider a simple, fundamental experiment. A sheet of iron about a quarter



CROSS-SECTION EXPERIMENT using slow neutrons from the reactor at Brookhaven National Laboratory is depicted in this photograph. Bursts of neutrons from the pile pass through the target material (*not shown*) and then into the long helium-filled balloon in the left foreground. A scintillation counter immediately behind the balloon detects the incoming neutrons, separating them as to energy by their time of arrival. The helium path is provided because nitrogen atoms in air absorb neutrons and would weaken the beam.



NEUTRON CROSS SECTIONS of all the elements as predicted by the cloudy-crystal-ball model are shown at the top. The actual values found by experiment appear below. The graphs are three-dimensional, with atomic weight increasing horizontally to the

right, the energy of the probing neutrons increasing in the direction from the plane of the page out toward the reader, and the cross section increasing vertically upward. The range of neutron energies shown extends from zero to about three million electron volts.

of an inch thick is exposed to a stream of neutrons flowing at a certain known rate, in particles per second, with all the particles moving at about the same speed. Some of the neutrons will pass straight through the thin sheet of iron. Some will be deflected by collisions with nuclei of iron atoms. Some will be captured, or absorbed, by nuclei. Behind the sheet a detector counts the number of neutrons that have passed straight through (those that are deflected miss the detector). Subtracting the number of neutrons that hit the detector from the number that entered the sheet, we get the number scattered or absorbed.

All these neutrons have collided with iron nuclei (assuming that the sheet is pure iron). Hence the percentage of neutrons scattered or absorbed tells us what fraction of the cross-sectional area

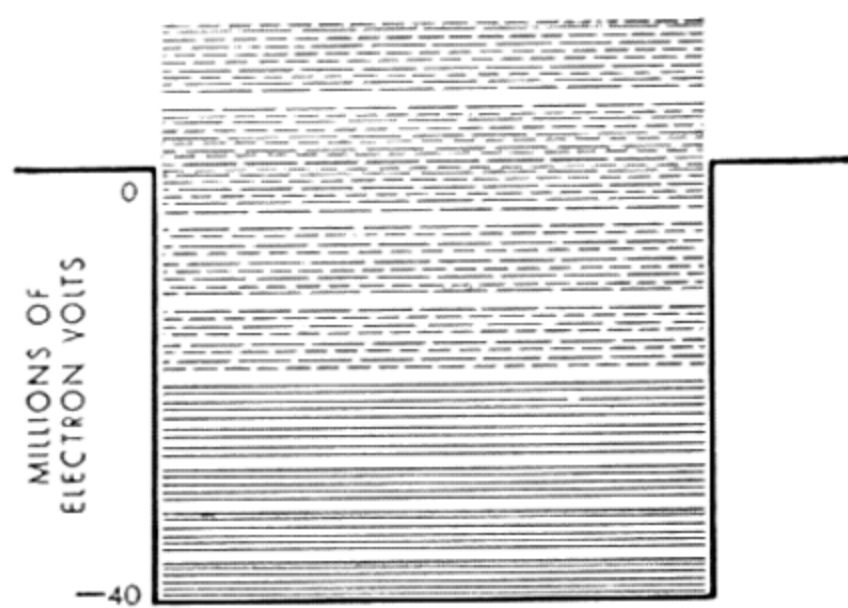
hit by the beam is effectively covered by nuclei. We know, from the weight of the iron, the number of atoms in this area. Thus we can calculate the area of interception represented by each nucleus. This area is called the scattering and absorption cross section of the iron nucleus for neutrons of the given speed.

When the experiment is performed with fairly energetic neutrons (about 15 million electron volts), the cross sections of various nuclei turn out to obey a rather simple law. Each nucleus has an effective area about equal to that of a circle whose radius is found by the formula: 1.4 times the cube root of the atomic weight times 10^{-13} centimeters. The length unit 10^{-13} centimeters has informally been given the name "fermi," after the late Enrico Fermi. The atomic weights of the natural elements range

from one to 238, so the cross-section radii range in length from 1.4 fermis to about 9 fermis.

Experiments with other types of bombarding particles, such as protons, yield the same result, when we make corrections for the effects of electrostatic force in their case. It appears that the cross section actually is a physical area.

Now we can describe the foregoing experiment in other terms, considering the neutron not as a particle but as a wave. We translate particle language into wave language by means of these two definitions: (1) the amplitude, or strength, of a wave at any point indicates the probability of finding a particle at that point—a strong wave means a high probability; (2) the frequency of a wave measures the energy of its parti-



WELL PICTURE of the nucleus with discrete energy levels is diagrammed schematically. Solid horizontal lines represent orbits which are normally filled; dotted lines are orbits which are not occupied unless a nucleon acquired more than its normal energy. Levels above the top of the well indicate that a nucleon may have more than 40 million electron volts of motion-energy and still revolve in an orbit within the nucleus.

cle—high frequency (short wavelength) means a fast particle, low frequency (long wavelength), a slow particle.

In wave terms the cross-section experiment might read as follows: A sheet of iron is placed in a monochromatic (single-frequency) beam of short-wave neutrons. (An energy of 15 Mev corresponds to a wavelength of about six fermis.) The scattering and absorption of neutrons is measured by the dimming of the brightness of the beam in its passage through the sheet. In these terms, for neutrons of the specified wavelength the iron nucleus acts as an opaque ball with a radius of 1.4 fermis. It blocks a portion of the neutron "light" beam and casts sharp shadows.

This wave description is very useful in interpreting the experiment. It suggests an optical model.

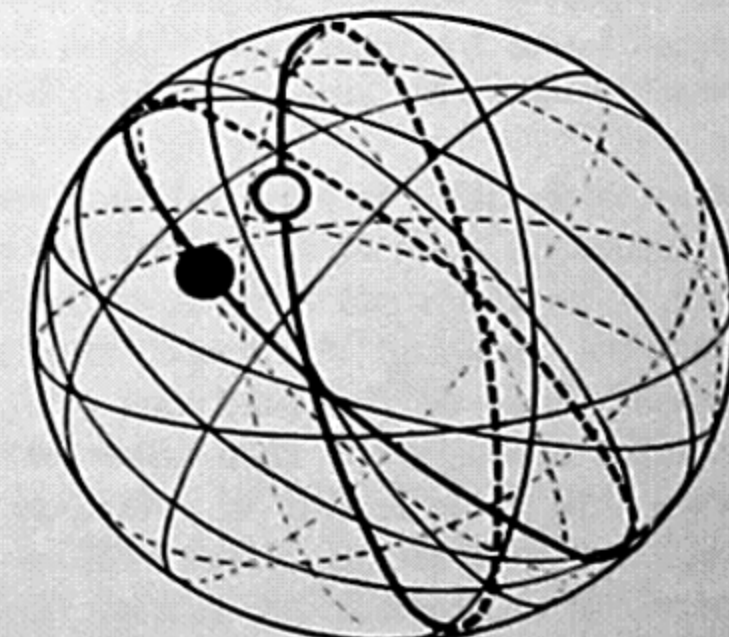
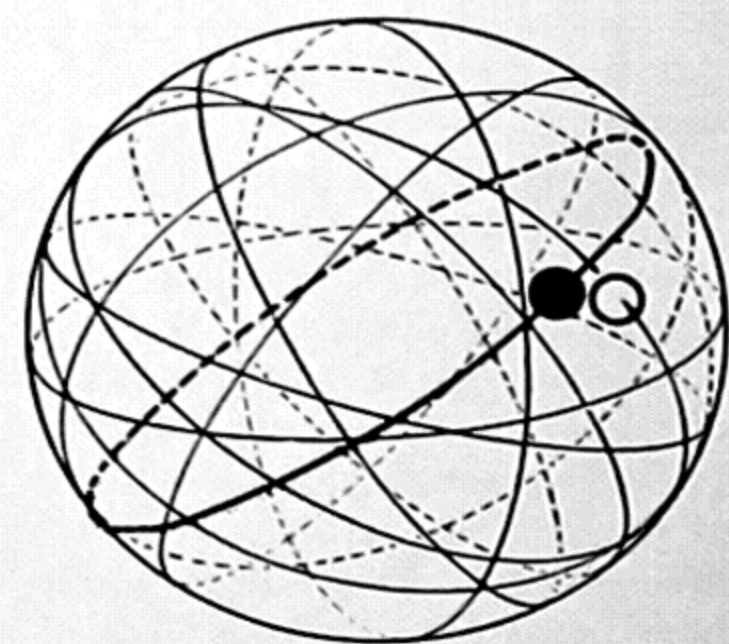
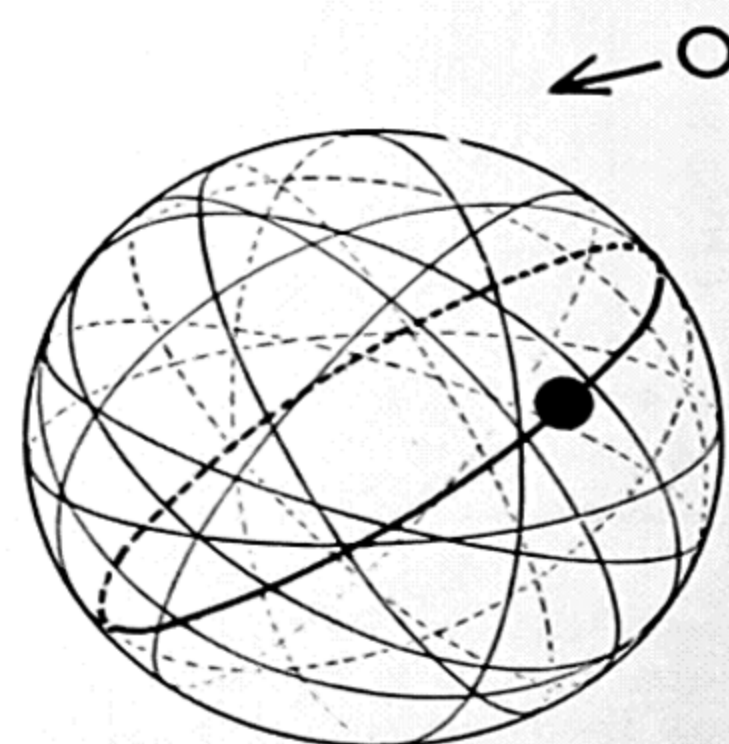
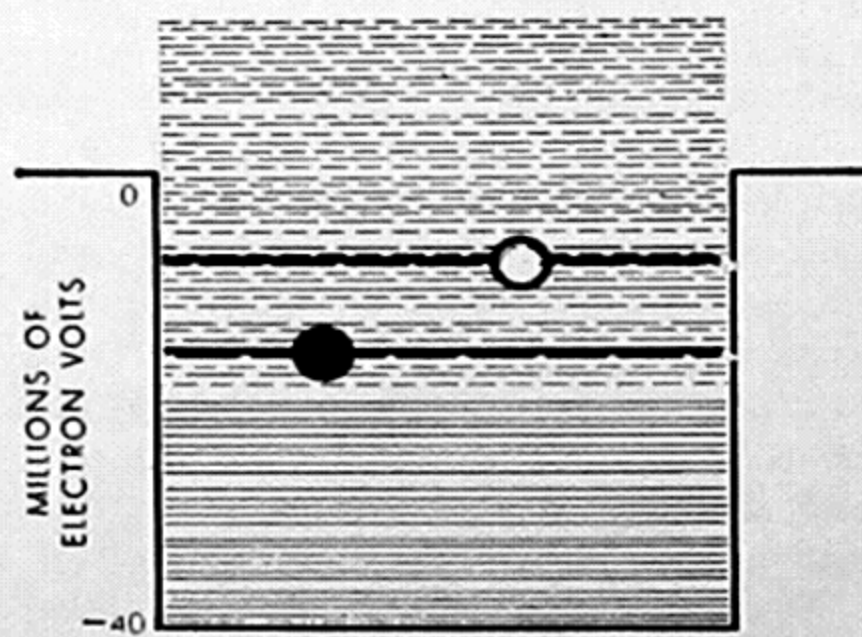
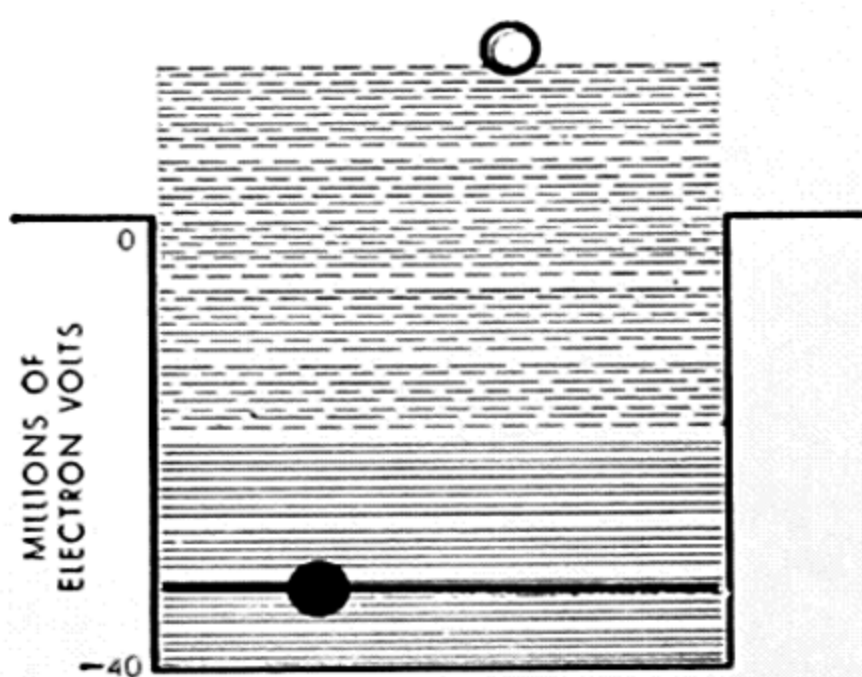
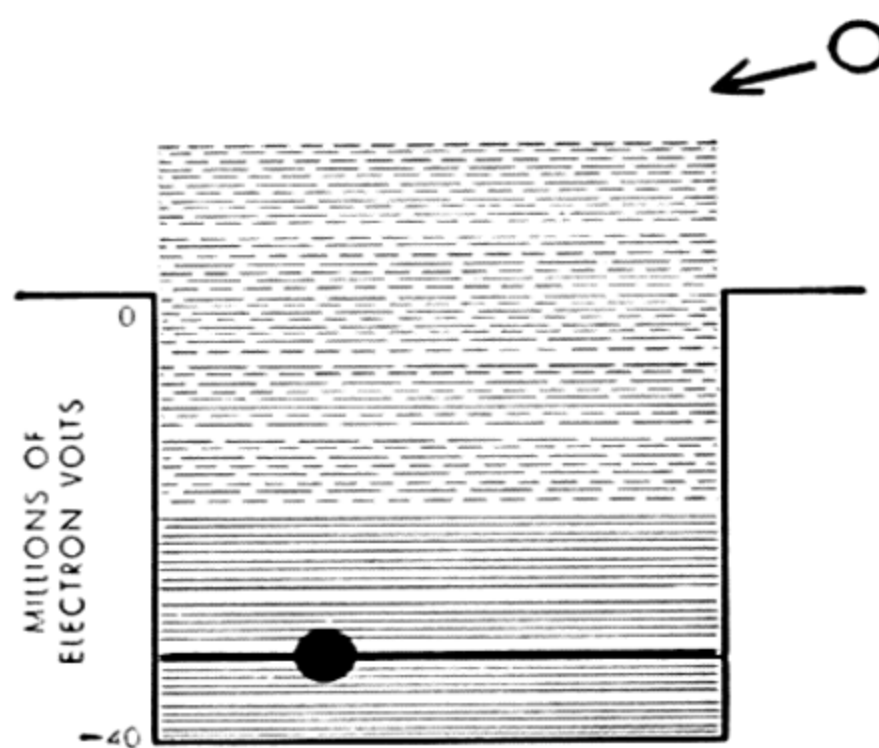
Thinking in optical terms, we are at once led to suspect that the results of cross-section experiments should depend on the wavelength of the neutrons used. For one thing, a beam whose wavelength is considerably longer than the nuclear radius will not cast sharp shadows, just as a long water wave washing around a small rock will not cast any appreciable "shadow" of calm water behind it. Yet the rock will cut off a little of the water wave's energy. If a harbor were dotted with separate small, exposed rocks, one could compute how much energy they would extract from incoming waves. In much the same way we can figure out how much energy an assembly of opaque nuclei should subtract from a long-wave neutron beam. When this computation is compared with actual measurements, however, it is

found to be far off. The nuclei do not block nearly as much energy as they would be expected to. In fact, they act as though they were almost transparent to the waves! Nuclear matter, regarded as optical material, is opaque to short-wave neutrons but becomes more and more transparent as the wavelength of the neutrons increases. As observed in

the "light" of slow neutrons, the nucleus looks like a cloudy crystal ball.

Let us see how this model helps us to understand the phenomena observed in experiments.

A beam of light passing through a group of semitransparent glass balls would, of course, be dimmed. Some



NEUTRON ABSORPTION is illustrated in these drawings. Left-hand diagrams show energies, right-hand diagrams, locations in space. At the top, a bombarding neutron (open dot) approaches the nucleus. Solid dot represents the nucleon destined to be hit, colored lines, its orbit. Middle diagrams show particles just before impact, when no energy has been exchanged. In the final result (bottom) the neutron has lost energy, dropping into a vacant orbit, and the target nucleon has picked up the energy, jumping to a higher orbit. The energy transaction is possible only if both particles find vacant positions. Cloudiness, or tendency to absorption, increases with the energy of the bombarding neutron because high-energy impacts put both particles into the upper, sparsely populated orbit region, while low-energy impacts leave them in the lower region where they are less likely to find vacancies.

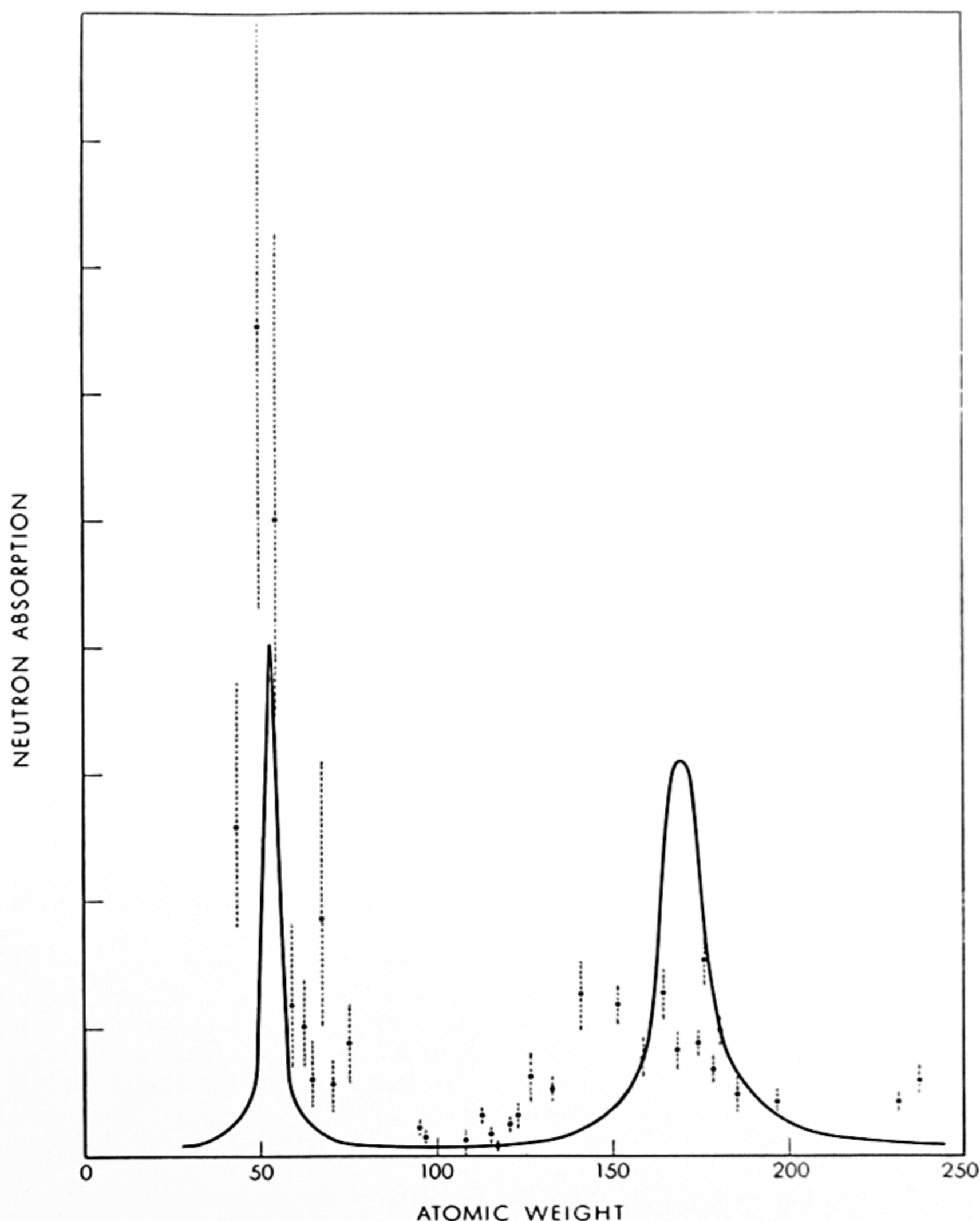
light would be lost because of bending of rays out of the beam, the amount depending on the index of refraction of the glass. Some would be lost by absorption within the balls, the amount depending on the glass's coefficient of absorption, or cloudiness. If the index of refraction and the absorption coefficient are known, we can calculate by optical theory the amount of dimming for balls of any size and spacing.

By means of certain experiments it is possible to estimate the refraction index and the absorption coefficient of nuclear matter. We can then predict how strongly a neutron beam of any given wavelength, or energy, will be dimmed by a group of nuclei of any given size. The results of such a calculation are shown in the upper graph on page 501. The panels with curved upper edges indicate the theoretical cross sections of the elements (indicated by atomic weight) for neutrons of various energies ranging from zero to about three Mev. Each panel represents an atom of a certain size, and the curve of the top edge plots the predicted variation of the nucleus' cross section with increasing energy of the neutron beam. The height of each point on the curve stands for the size of the total cross section (scattering and absorption) for that energy.

The pattern is not simple. Usually the cross section decreases with increasing energy (shorter wavelength) of the neutron "light," but in some cases it rises along part of the range and there is a hump in the middle of the curve. And as we look along the sequence of the elements, we can see that there are groups of elements, around atomic weight 40 and again between 100 and 140, whose cross section for low-energy neutrons is small rather than large.

If the actual measurements of cross sections were to match this intricate and peculiar theoretical pattern, we could certainly feel that it was no accident. The measurements have been made (by many independent experimenters), the curves have been plotted, and the agreement between the pattern based on the measurements and the pattern based on the theoretical predictions is astonishing. The results of the measurements are shown in the lower graph on page 501. These curves match almost exactly many of the detailed and complex variations in the theoretical curves.

Evidently the cloudy crystal ball model is a highly useful way of picturing nuclear matter. But the model suggests some new puzzles. To say that the nucleus is semitransparent implies that part of the wave energy that passes



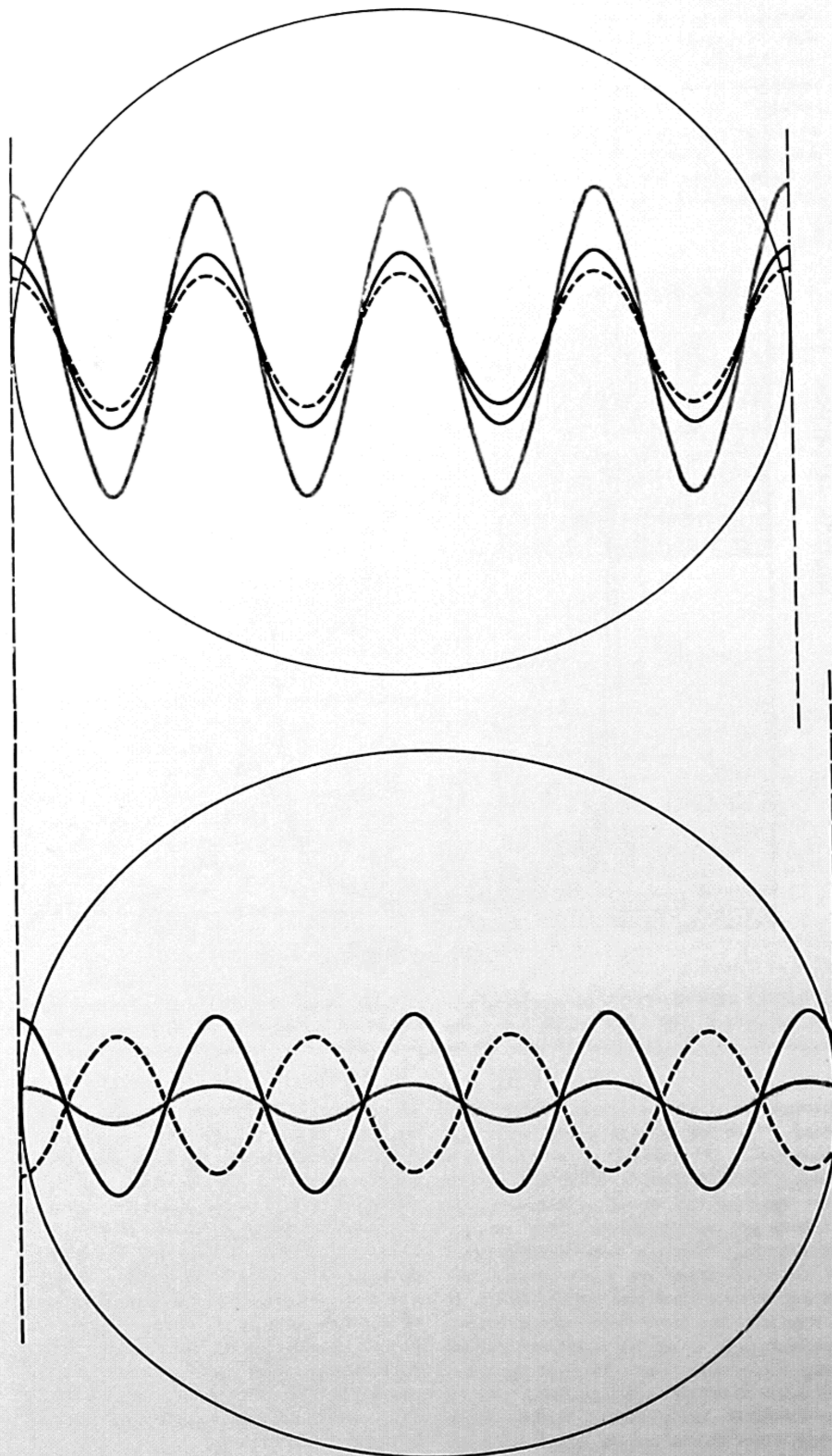
STRONG ABSORPTION for nuclei of atomic weights near 55 and 155 is predicted by the optical model. The solid curve shows the absorption called for by the theory, the dots represent experimental observations. Dashed lines indicate the limits of experimental error.

through it emerges intrinsically unaltered. The emerging wave may be changed in direction but not in wavelength. Now in particle terms this means that some of the neutrons entering the nucleus get out again with their energy unchanged, which is surprising indeed in view of what we know about the strong interactions among nucleons. It is true that the force fields of the nucleons act only over an extremely short range—less than three or four fermis—but a bombarding neutron that gets into the nucleus surely comes within these fields. How can it pass through without interacting with the particles in the nucleus and exchanging energy?

For help in resolving this paradox we can turn to the shell model of the nu-

cleus [see "The Structure of the Nucleus," by Maria G. Mayer; *SCIENTIFIC AMERICAN* Offprint 228]. In brief, this model says that the protons and neutrons inside the nucleus are arranged in a system of orbits, or shells, as the electrons outside the nucleus are. Each orbit corresponds to a specific level of energy, and a nuclear particle can pass from one to another only by an abrupt "quantum jump." Furthermore, the famous Pauli exclusion principle applies to the nucleons as to electrons. Each nuclear orbit can contain only two particles, spinning in opposite directions.

Now these occupancy rules imply "social" behavior on the part of the nucleons. Any individual particle occupies the orbit that it does because others are



TOTAL WAVE MOTION inside nucleus is strong (gray curve in top diagram) when incoming wave (solid black curve) and reflected wave (broken curve) reinforce each other. If nuclear radius is wrong length (bottom) individual waves interfere giving small total.

filled. In some sense it "knows" the over-all situation and stays in its proper niche. To put it another way, the force on an individual nucleon is a group force stemming from the entire ensemble of its neighbors acting as a single body. So, also, neutrons entering the nucleus from the outside are confronted with a single force exerted by the entire population of resident nucleons. Many of the incoming neutrons respond only to this force, failing to interact with individual nucleons. It is as if an individual entered a room full of dancers. He might try to cross the floor, in which case he would bump into some of the dancers or perhaps even be grabbed by a partner. But on the other hand he might skirt the group and emerge from the other side of the room without having lost or gained any momentum. So a neutron, swung away from individual contacts by the group force of the nucleus, may pass through it as if it were transparent.

Now to complete the picture of the energy conditions within the nucleus we shall forget cloudy crystal balls and shells for a moment and resort to still another model. In considering the energy states of the nucleons, it is convenient to think of them as having negative energy. That is, we arbitrarily choose zero as the top of the energy scale within the nucleus and represent the lower energy states by negative numbers, just as we can speak of temperatures below zero. In the nuclear case the reason for doing this is to indicate that energy must be added to a nucleon to pull it free of the other nucleons' force fields and raise it to the isolated, motionless state of zero energy. The nucleons are analogous to the water in a well, which has negative potential energy, with respect to ground level, because work must be done to raise the water to that level. We take a well, then, as a model of the nucleus.

It is a "potential" well—a well simply representing the fact that various levels of negative potential energy exist in the nucleus. These levels correspond to the orbits that can be occupied by the nucleons [see diagram on page 502]. Note that the well is "square," with sharp corners at the top, i.e., no sloping edge. This reflects an assumption that a particle just outside the nucleus feels no force while one just inside feels the full amount. A particle is either completely in or completely out of the well. As for the well's depth, experiments indicate that it takes about 40 Mev of energy to lift a particle out of the nucleus from the lowest level.

This well model is a precise counterpart of the cloudy crystal ball. The depth of the well is the measure of the nucleus' index of refraction, or the crystal ball's effect as a single body upon a neutron passing through it—an effect which may change the neutron's direction but not its energy. When we say that the crystal ball is transparent to the neutron, it is the same as saying that the neutron has not fallen into the well, because it has been acted on by the potential (single-body force) of the well as a whole. When we say that the ball is cloudy, we mean there is a finite probability that the neutron will be absorbed by the nucleus, or, in terms of energy, will fall into the well (exchange energy with individual nucleons). This probability is the coefficient of absorption. In other words, the cloudiness of the nucleus is determined only by absorption, not by scattering.

The crystal-ball concept pictures the nucleus in terms of waves; the potential-well model considers it in terms of particles. Both pictures are consistent with and supplement each other, when we examine the energy transactions in the nucleus.

Let us return to the cloudy crystal ball and try to discover what further investigations it may suggest.

We have observed that the cloudiness (absorption) of the crystal ball depends on the energy of the neutron "light" that attempts to pass through it. The lower the energy of the neutron beam, the more transparent the nucleus becomes. But this is not the whole story. Remembering that the neutrons behave as waves, we realize that, from a theoretical point of view, nuclei of certain sizes

should be opaque even to neutrons of low energy. We should expect a "resonance" effect to operate under certain circumstances. These circumstances are the creation of a "standing wave" within the nucleus. Every physics student knows that when sunlight passes through a water drop, part of its energy is reflected back into the drop from the far boundary. Such reflection occurs whenever waves encounter a boundary between one medium and another, and the boundary of an atomic nucleus should be no exception. Now if a drop (or nucleus) is just the right size to allow the incoming waves and the reflected waves to reinforce each other, they will form what is known as a standing wave. This contains so much energy that even a poor absorber will soak up a considerable amount of it. Thus a nucleus in which a standing wave is set up should absorb strongly a low-energy wavelength to which most other nuclei are transparent.

We know what the size of the drop (or nucleus) must be in relation to the wavelength in order to have a standing wave. It has been calculated that a standing wave will occur whenever the radius of the body is an odd multiple of the wavelength divided by four. We know that the wavelength of the neutrons entering a nucleus is 4.4 fermis. Dividing by four, we get 1.1 fermis. Any nucleus whose radius is an odd multiple of this distance—*e.g.*, 3.3, 5.5, 7.7—should exhibit strong absorption of neutrons of the 4.4-fermi wavelength; in other words, it should be cloudy.

The elements with these radii are those of atomic weight 11, 55 and 155. Several experimenters have tested the prediction by measuring the cross sec-

tions of nuclei in the neighborhood of those atomic weights, and they have found that the 55 nucleus definitely is cloudy to the slow neutrons; that 11 probably is, and that nuclei near 155 also show the predicted absorption effect when corrections are made for the fact that their shape is not spherical.

Thus the cloudy crystal ball model not only accounts surprisingly well for a number of previously known experimental results but has also predicted new ones. It must reflect an important aspect of the real nucleus.

Many theoretical physicists are now trying to refine and extend the model. One problem is to determine whether the nucleus has a sharply defined surface as far as its optical properties are concerned, or whether it begins thinly and gradually becomes denser toward the center. The square-well concept assumes that it has a sharp boundary. By rounding off the corners of the well—*i.e.*, varying the values for the index of refraction near the surface—one might derive cross-section curves which would agree with the measurements even more closely than the theoretical set shown here.

It should be emphasized that the cloudy crystal ball represents only gross properties of nuclear matter. Calculations based on the model yield only average values. The model cannot account for the detailed variations in a nucleus' cross-section curves—the fine structure of ups and downs that appears in a chart of actual measurements. Yet despite its coarseness, in fact, because of it, the optical model can be expected to yield further valuable insights into the ultimate nature of matter.

The Authors

VICTOR F. WEISSKOPF and E. P. ROSENBAUM are, respectively, a theoretical physicist and a science editor. Weisskopf, who was born in Vienna, studied with Wolfgang Pauli in Zürich and with James Franck and Max Born at the University of Göttingen, where he took his Ph.D. in physics in 1931. He spent the middle 1930s in research in Europe, notably at the Universities of Copenhagen and Cambridge. At Copenhagen he studied with Niels Bohr. On his arrival in the U. S. in 1937 he first taught at the University of Rochester. From 1943 to 1947 he was a theoretical physicist at the Los Alamos Scientific Laboratory. He was appointed professor of physics at the Massachusetts Institute of Technology in 1945. Rosenbaum is a member of the board of editors of SCIENTIFIC AMERICAN. He was born in New Haven, Conn., in 1916, attended both

Harvard and Yale as an undergraduate and graduated from Yale in 1937. He then taught mathematics and physics at the Milford School, a preparatory school in Connecticut. During World War II, as a captain in the Air Force, he served at Wright Field as project engineer on the first Air Force radar-directed guided missile. After the war he returned to teaching and administrative work at the Milford School, but then turned to writing, principally for radio and television. He joined the staff of SCIENTIFIC AMERICAN in 1952.

Bibliography

- INTRODUCTORY NUCLEAR PHYSICS. David Halliday. John Wiley & Sons, 1955.
- A MODEL FOR NUCLEAR REACTIONS WITH NEUTRONS. H. Feshbach, C. E. Porter and V. F. Weisskopf in *The Physical Review*, Vol. 96, No. 2, pages 448-464; October 15, 1954.
- NIELS BOHR AND THE DEVELOPMENT OF PHYSICS. Edited by W. Pauli. McGraw-Hill Book Company, Inc., 1955.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

WATER

by Arthur M. Buswell and Worth H. Rodebush

Although we take its properties for granted, they are most unusual. As an example, it has the rare property of being lighter as a solid than as a liquid. If it were not, lakes would freeze from the bottom up.

Water is the only common liquid on our planet. Next to air it is the substance with which we are most intimately acquainted. Because it is so familiar, we are apt to overlook the fact that water is an altogether peculiar substance. Its properties and behavior are quite unlike those of any other liquid. To take just one example, water has the rare property of being denser as a liquid than as a solid, and it is probably the only substance that attains its greatest density at a few degrees above the freezing point (four degrees centigrade). The consequences of this behavior are of great importance to life on our planet. When ice forms on a lake, for example, its lower density (only nine tenths that of liquid water) keeps it on top and it acts as an insulating blanket to retard cooling of the underlying water. As a result lakes in the temperate zones do not freeze solid to the bottom but leave a zone for the winter survival of aquatic life. On the other hand, the same peculiar property of water has fatal consequences for living cells. When water in the cells freezes and becomes less dense, its expansion damages or breaks up the cells.

Even the elements of which water is composed—oxygen and hydrogen—are chemically exceptional. Both are unusually reactive. Oxygen is our chief source of energy, being responsible for the respiration of living organisms and the combustion of fuels. Hydrogen, unique in the fact that it has no enclosing shell but only a single electron, is able to attach itself to other atoms not only by means of its electron (a valence bond) but also by virtue of the attraction of its unoccupied, positively charged side for an electron in a second atom. This attachment is known as the hydrogen

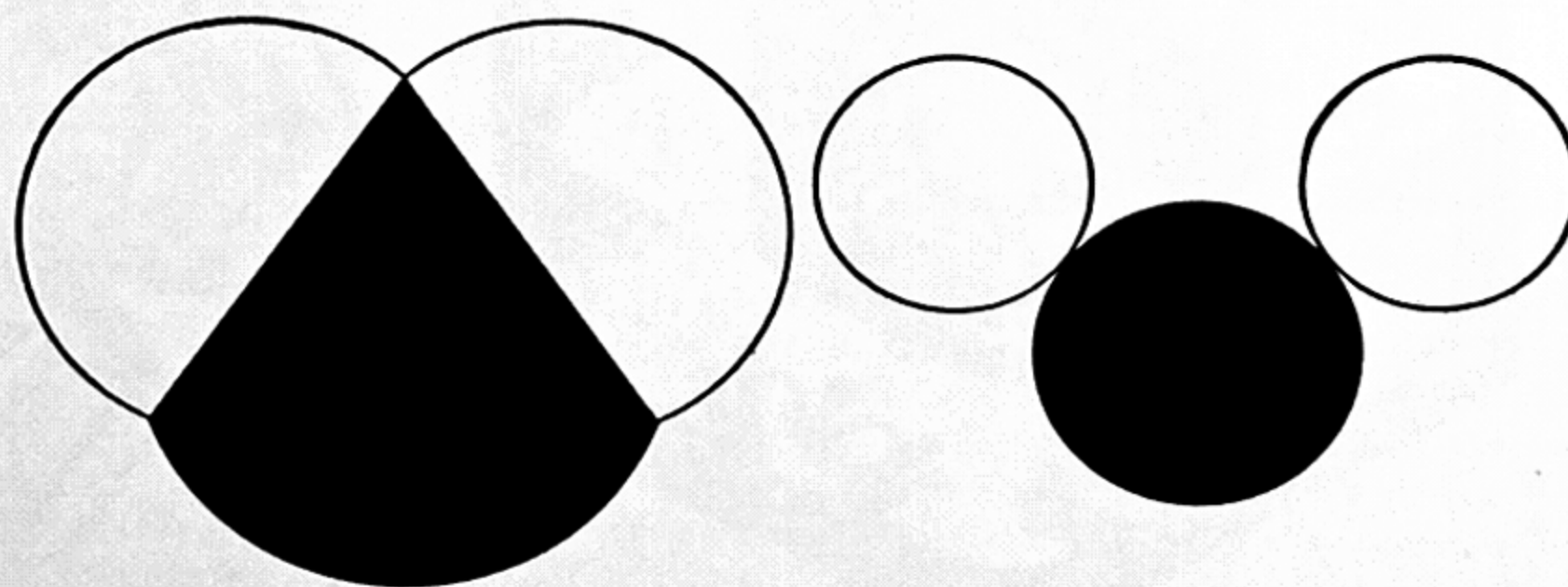
bond. In water the two hydrogen atoms attached to each oxygen atom can become linked to other atoms as well by means of these so-called hydrogen bonds. As a consequence the H_2O molecules are joined together, so that water should be considered not a collection of separate molecules but a united association. In effect the whole mass of water in a vessel is a single molecule.

The best method of detecting hydrogen bonds is to study water with the infrared spectrograph. We have found that the hydrogen bond absorbs radiation most strongly at a wavelength of about three microns, which is in the near infrared region of heat radiation—i.e., close to the visible light spectrum. Liquid water absorbs this radiation so powerfully that if our eyes were sensitive to the infrared, water would appear jet black. There is some absorption even in the visible spectrum at the red end. The fact

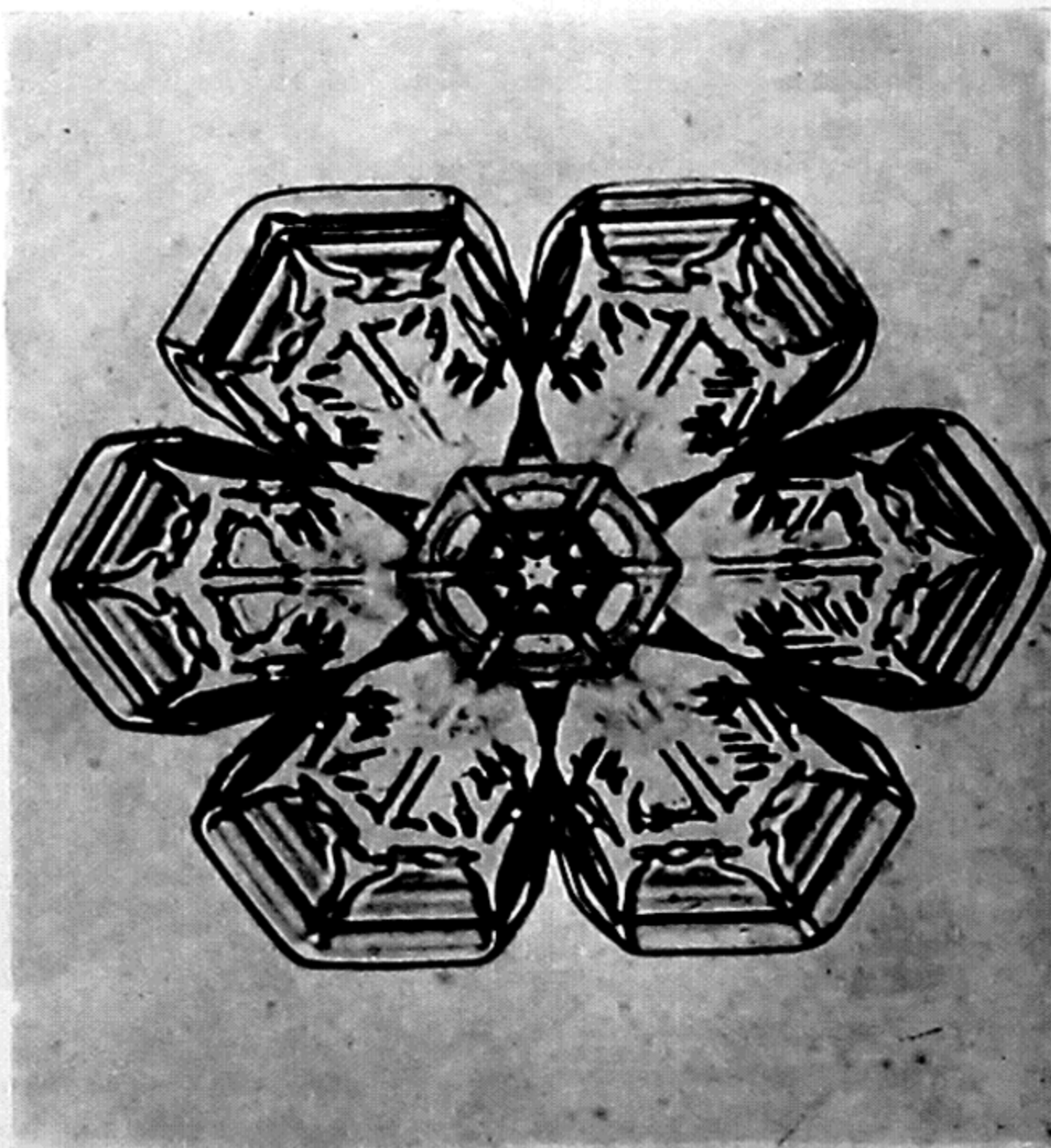
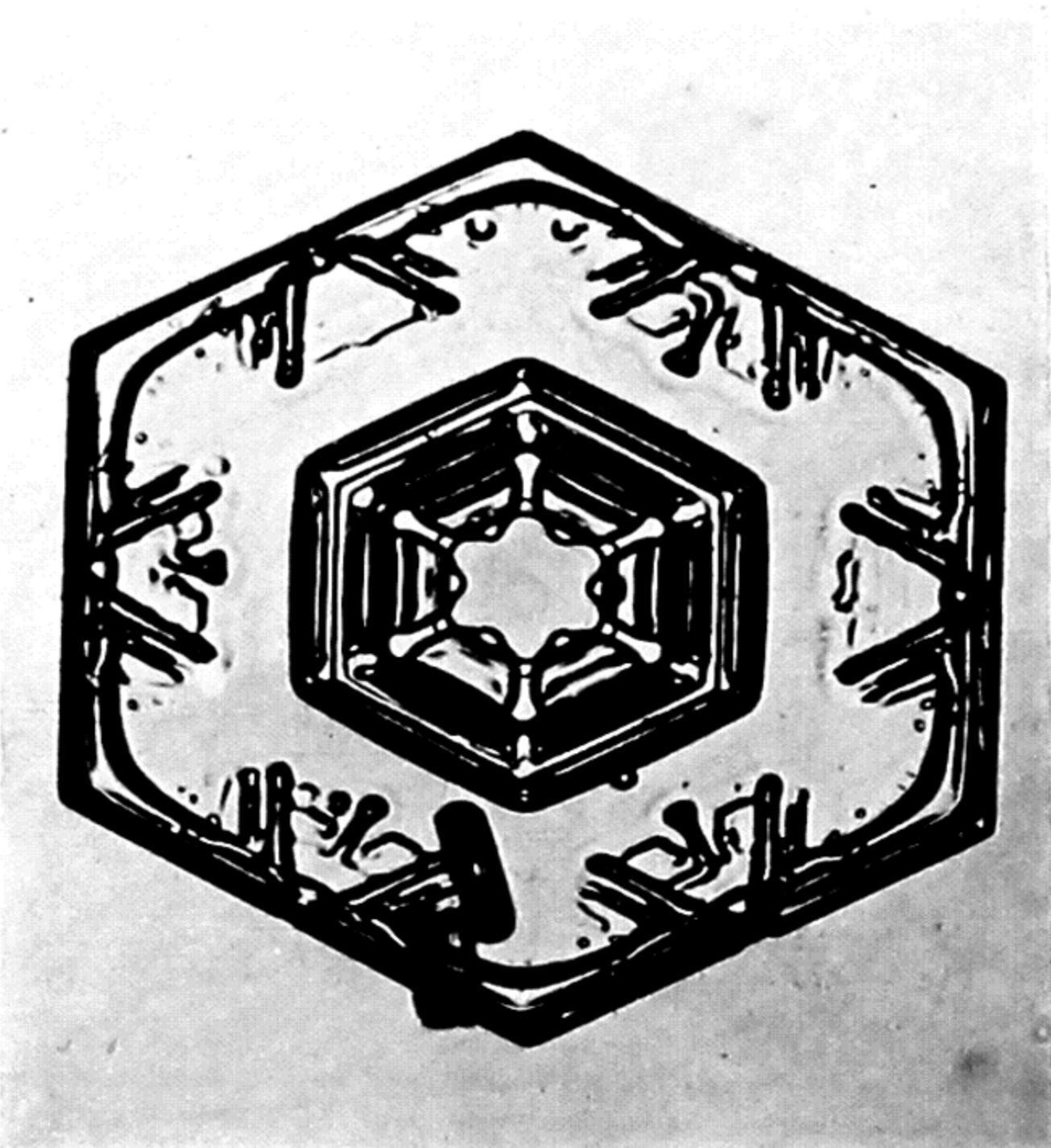
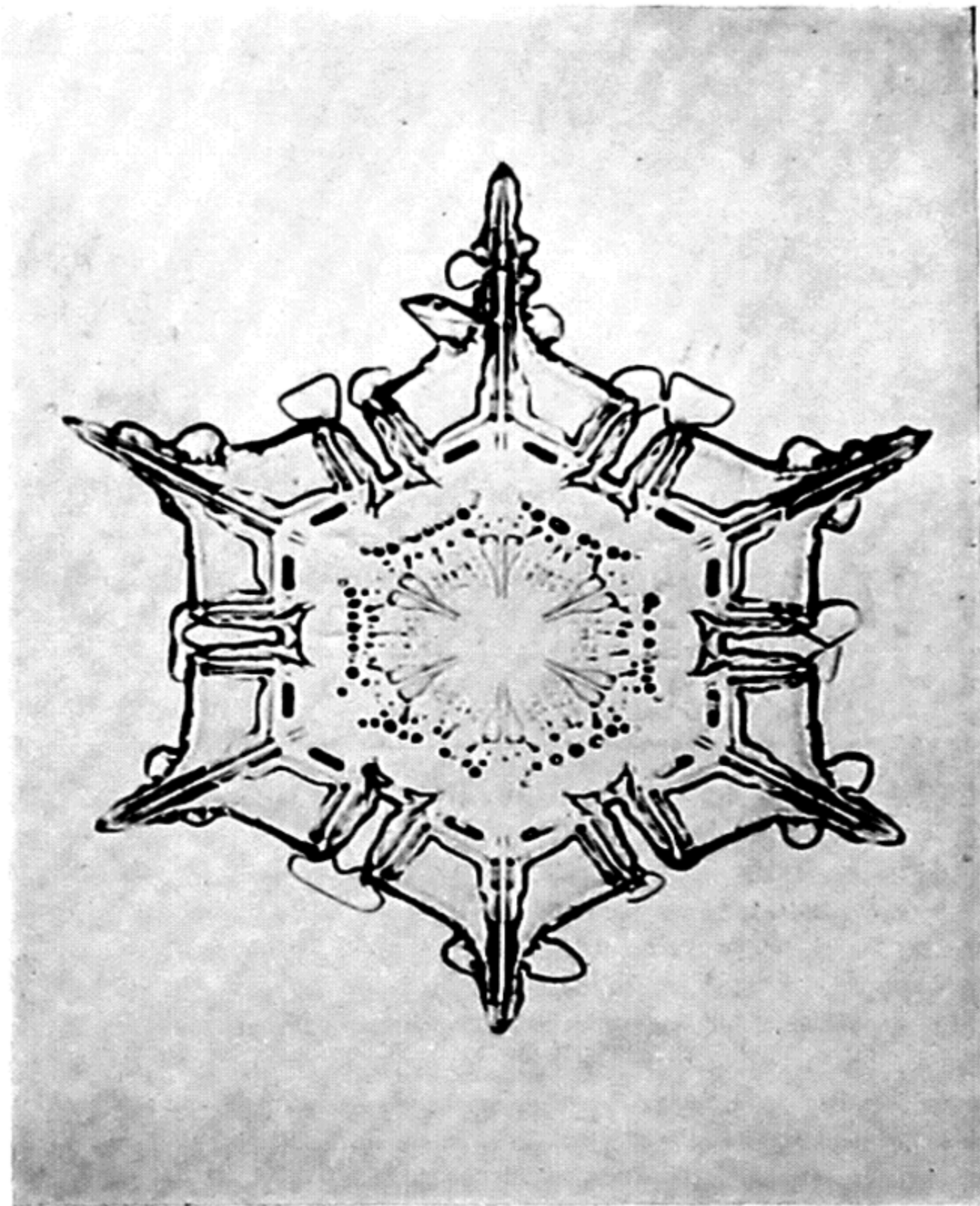
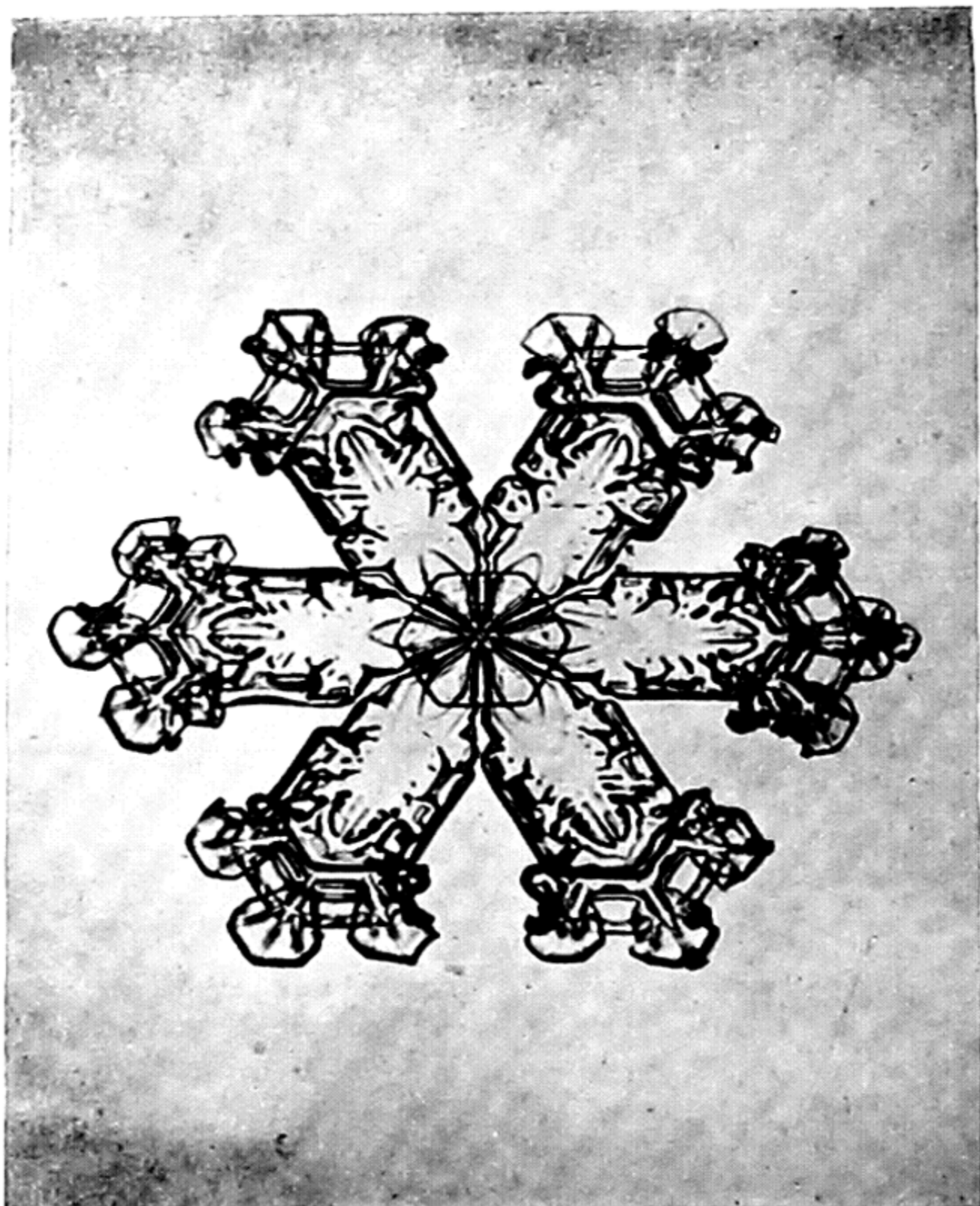
that water absorbs red light accounts for its characteristic blue-green color.

Water's 33 Components

One of the shocks to our familiar notions about water is that its formula is not simply H_2O . Nor is it a single substance. The beginning of this disillusionment came in 1934 when Harold Urey discovered "heavy water." Urey found that the purest water contained besides hydrogen and oxygen another substance like hydrogen but with an atomic weight of two, or twice that of hydrogen. This substance, which is now called deuterium, combines with oxygen to form the compound D_2O . By now, we know, of course, that there is a third isotope of hydrogen, called tritium, and three isotopes of oxygen: 0-16, 0-17 and 0-18. Thus the purest water that can be prepared in the laboratory is made up

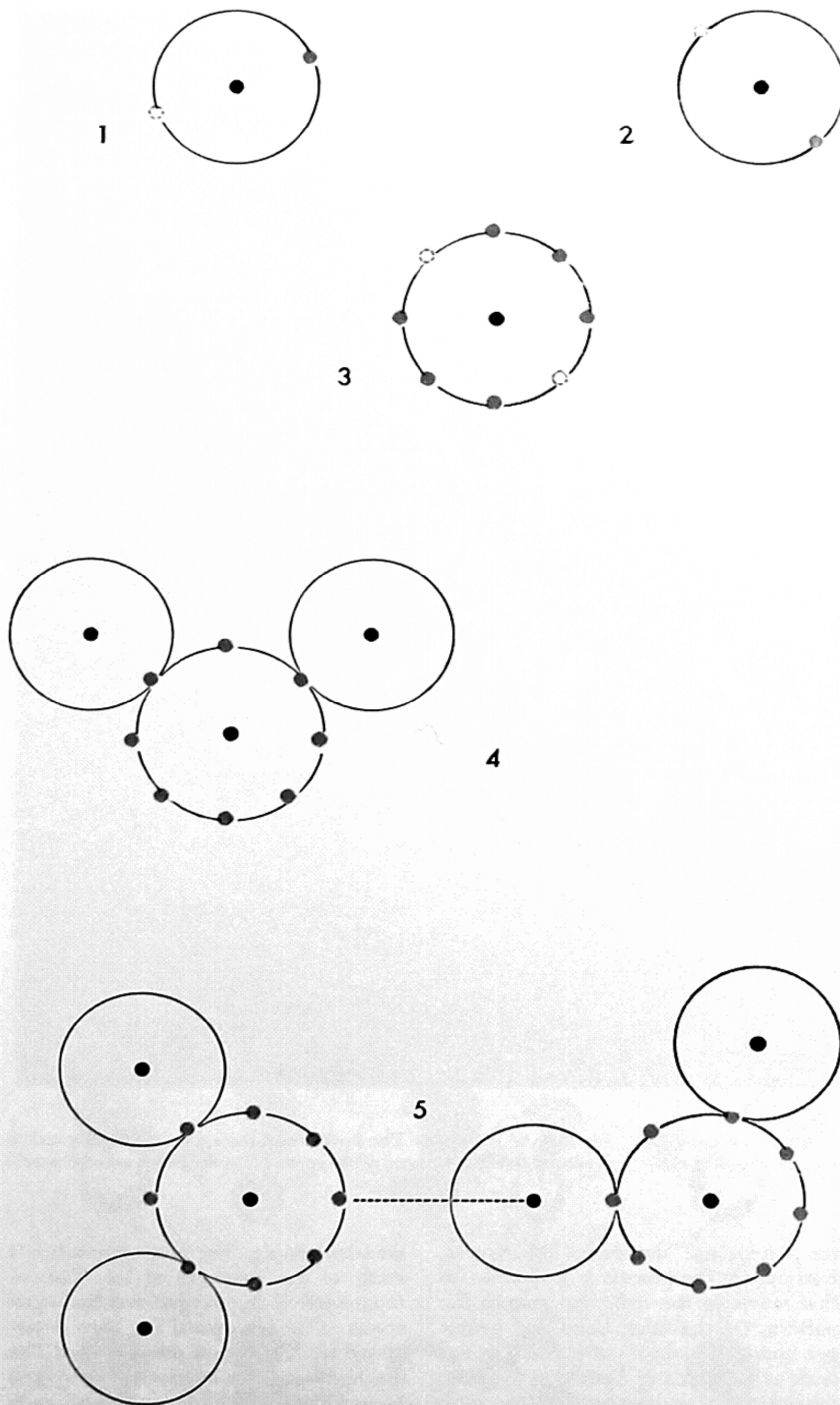


MOLECULE of water consists of one oxygen atom (*black*) and two hydrogen atoms (*white*). The distance between the center of the oxygen atoms and the center of each of the hydrogen atoms is .9 Angstrom unit (one Angstrom unit: .00000001 centimeter). The angle formed by the two hydrogen atoms is 105 degrees. These dimensions are fitted together in the drawing at left. In the more schematic drawing at right the size of the atoms has been reduced. This representation of the molecule is used in the following drawings.



SNOW CRYSTALS are enlarged about 50 diameters in these photomicrographs by Vincent J. Schaefer of the Munitalp Foundation

in Schenectady, N. Y. The hexagonal symmetry of the crystals is due to their molecular structure (*see diagram on page 511*).



ELECTRONS AND PROTONS of the water molecule account for most of its physical and chemical properties. The hydrogen atom (1 and 2 in this highly schematic picture) consists of a positively charged proton (black dot) and a negatively charged electron (colored dot). The oxygen atom (3) has eight electrons, six of which are arranged in an outer shell. Because hydrogen shell has room for one more electron (broken colored circle), and the outer shell of oxygen has room for two more electrons, the atoms have an affinity for each other. In the water molecule (4) the electrons of the hydrogen atoms are shared by the oxygen atom. Because the positively charged proton of the hydrogen atom now sticks out from the water molecule, it has an attraction for the negatively charged electrons of a neighboring water molecule (5). This relatively weak force (broken line) is called a hydrogen bond.

of six different isotopes, which may be combined in 18 different ways. If we add the various kinds of ions into which the addition or removal of an electron may transform water's atoms, we find that pure water contains no fewer than 33 substances [see top of page 512].

Of course the amounts of the isotopes other than common hydrogen and common oxygen (0-16) are tiny. Tritium and oxygen 17 appear only in the minutest traces, and deuterium is present to the extent of about 200 parts per million and oxygen 18 about 1,000 parts per million. However, the properties of heavy water, particularly the D_2O variety, have attracted wide interest and have been extensively studied.

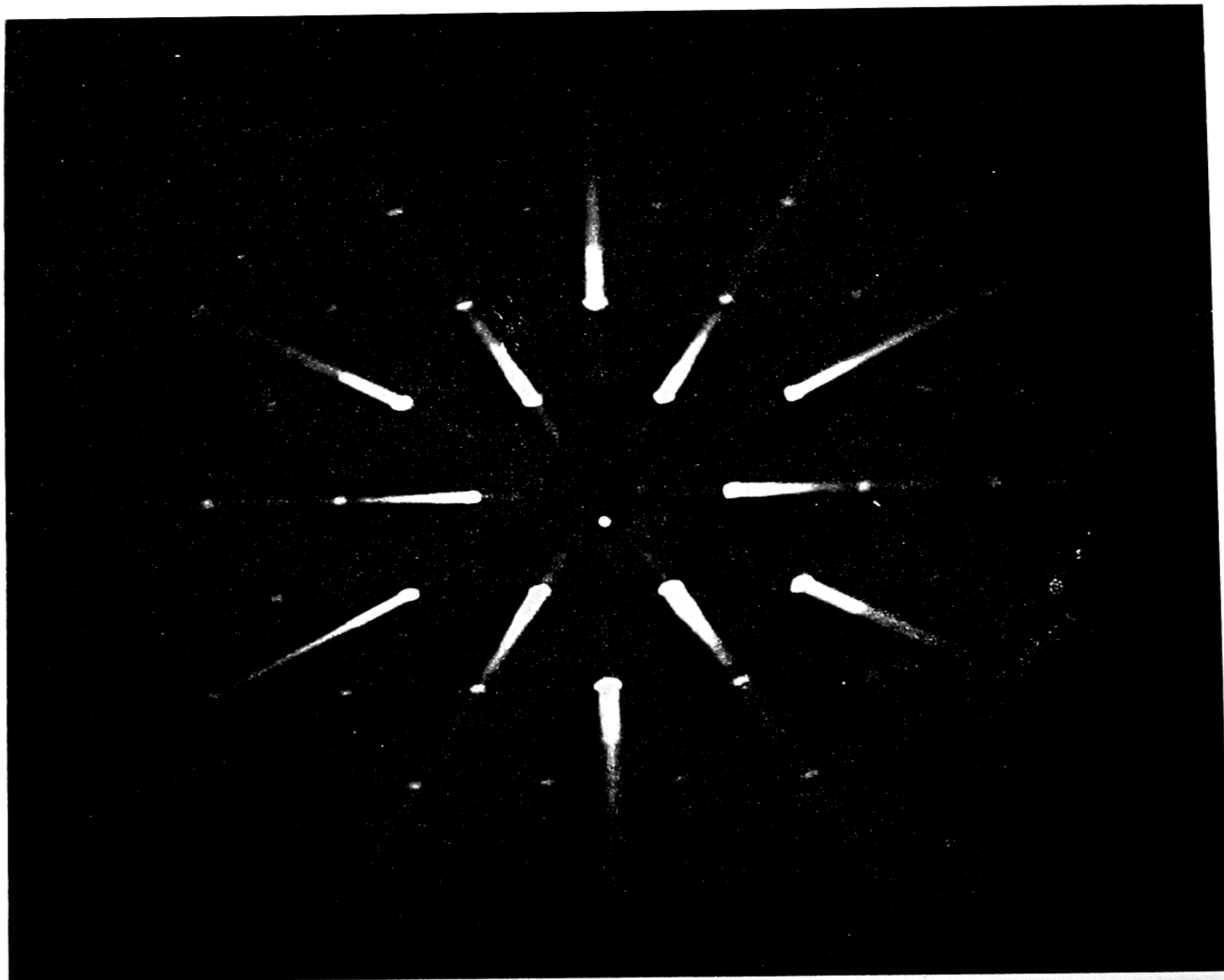
D_2O has a slightly higher boiling point than H_2O (101.4 degrees C.), freezes at a substantially higher temperature (3.8 degrees C.), and is somewhat more viscous than ordinary water. Its physiological properties are surprising. In animals and plants it appears to be entirely inert and useless. Seeds will not sprout in D_2O , and rats given only D_2O to drink will die of thirst.

The largest use of heavy water is as a moderator in nuclear reactors, but it is also widely employed in theoretical research, especially in organic and biological chemistry. If compounds containing active hydrogen are treated with D_2O , deuterium will replace the hydrogen and the compound will show changes in chemical properties resulting from the lesser reactivity of deuterium.

It is interesting to find that the amount of D_2O in natural water appears to be the same whether the water comes from an alpine glacier or the bottom of the ocean, from willow wood or mahogany.

Tritium is more ephemeral and more variably distributed. It is formed in the highest layers of the atmosphere by the bombardment of cosmic rays, and falls in rain and snow [see "Tritium in Nature," by Willard F. Libby; *SCIENTIFIC AMERICAN*, April, 1954]. Since tritium is radioactive, with a half-life of 12.5 years, it disappears after a time from water which has been out of contact with the atmosphere. Wines, and water in wells, can be dated by their tritium content. An interesting well in the Urbana-Champaign area was found to be devoid of detectable tritium, which means that at least 50 years have elapsed since the water fell as rain.

The functions of water in nature are innumerable. It is the solvent par excellence. It is the medium in which life originated and in which all organisms still exist. The living cell consists largely



X-RAY DIFFRACTION photograph of ice was made with a precession camera by I. Fankuchen and his colleagues at the Polytechnic

Institute of Brooklyn. The position of the spots in the photograph is related to the symmetry of the crystal (see diagrams on next page).

of water and literally floats in water. Considering how predominantly living matter is made up of this fluid, the extent to which it takes on a solid shape is surprising indeed.

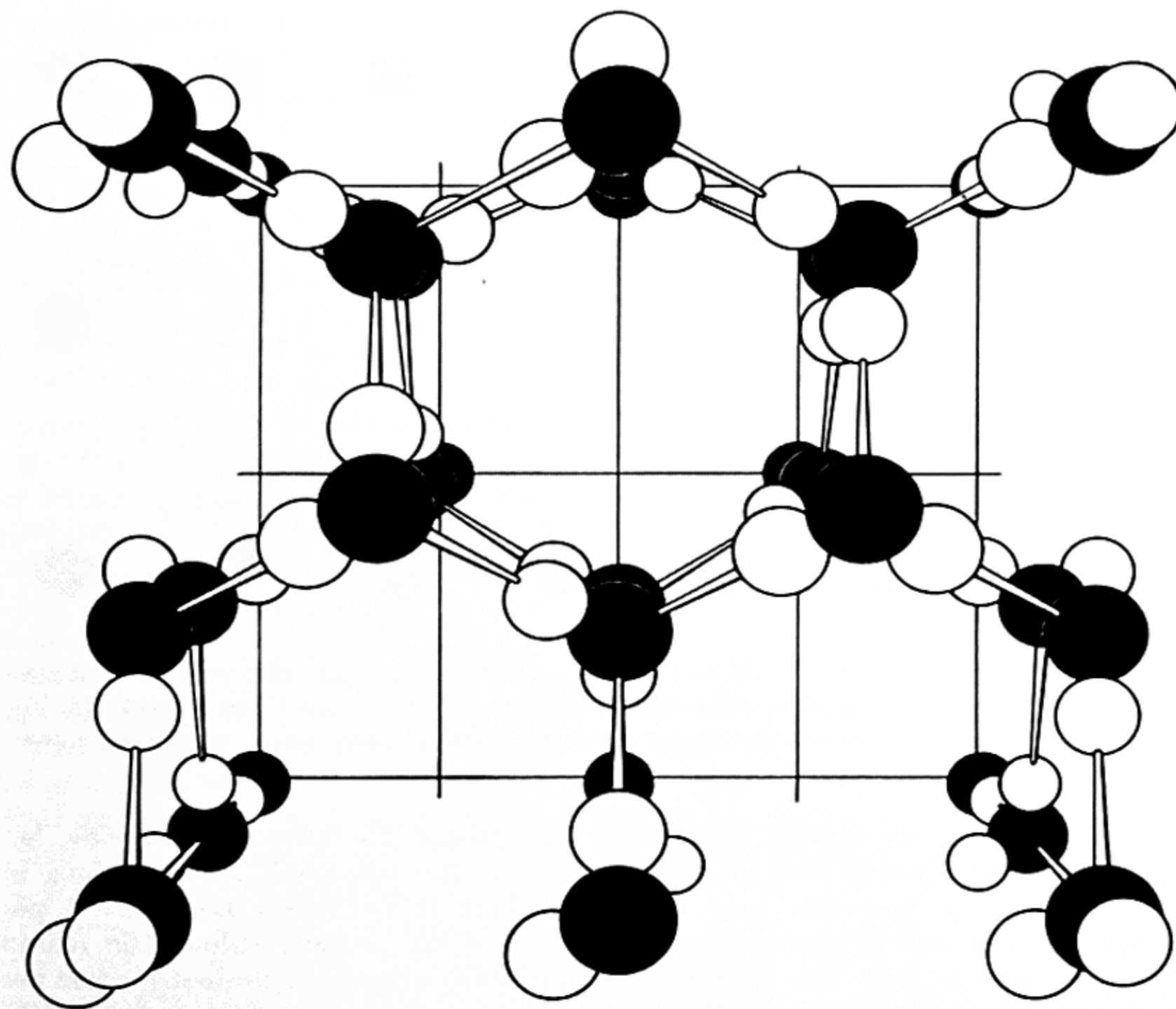
Water plays a fundamental role in the protein molecule, the basic material of living matter. Proteins have a structure which places them in the class of substances known to chemists as plastics. In order that a plastic may possess flexibility and other desirable physical properties, it must contain a fluid called a plasticizer. Water is a "plasticizer" for proteins. In the chainlike protein molecule the hydrogen bonds of water provide secondary links which fix the pattern of the molecule. Removal of water alters

the pattern and "denatures" the protein. Fortunately the process is reversible, so that water in the cells can restore the pattern. On the other hand, the hydrogen bonds of hydrides other than water, such as ammonia or hydrogen cyanide, form a stable denatured configuration which freezes the protein in a dead pattern. This is why all the hydrides except water are extremely toxic. Their action is somewhat like that of a virus, which corrupts the true protein structure into a strange distorted pattern.

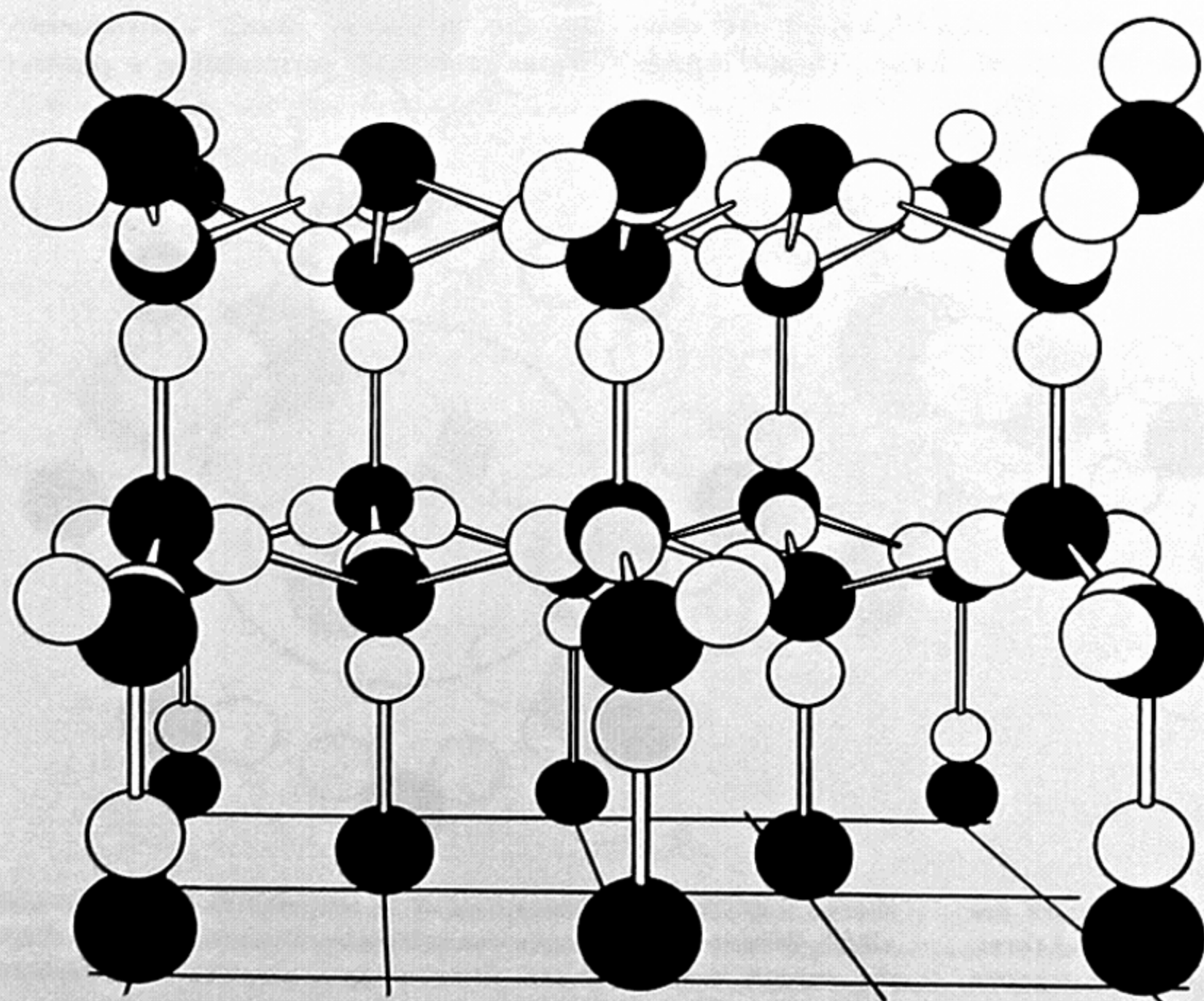
Water's Structure

To understand the behavior of water we must understand its structure. It is

far from simple. The best approach is a study of the structure of ice. The arrangement of the oxygen and hydrogen atoms in the ice crystal has been determined by X-rays and other means. The two hydrogens are bonded to the oxygen approximately at right angles to each other, more exactly, at an angle of 105 degrees [see illustrations on page 507]. If the angle were 109 degrees, the frozen water molecules would form a cubic lattice, as in the diamond crystal. But in this case such a structure would be unstable because of the strain on the distorted bonds. The exact arrangement of the molecules in the ice crystal is not known with certainty; we know that they form a hexagonal structure, which



ICE consists of water molecules in this arrangement. The top drawing shows a model of ice seen from one direction. The bottom drawing shows the same model seen as if the reader had turned the top drawing forward on a horizontal axis in the plane of the page. Some hydrogens have been omitted from the molecules which touch the grid. Each hydrogen in each molecule is joined to an oxygen in a neighboring molecule by a hydrogen bond (rods). In actuality the molecules of ice are packed more closely together; here they have been pulled apart to show the structure. In a similar model of liquid water the molecules would be much more loosely organized, farther apart and joined by more hydrogen bonds.

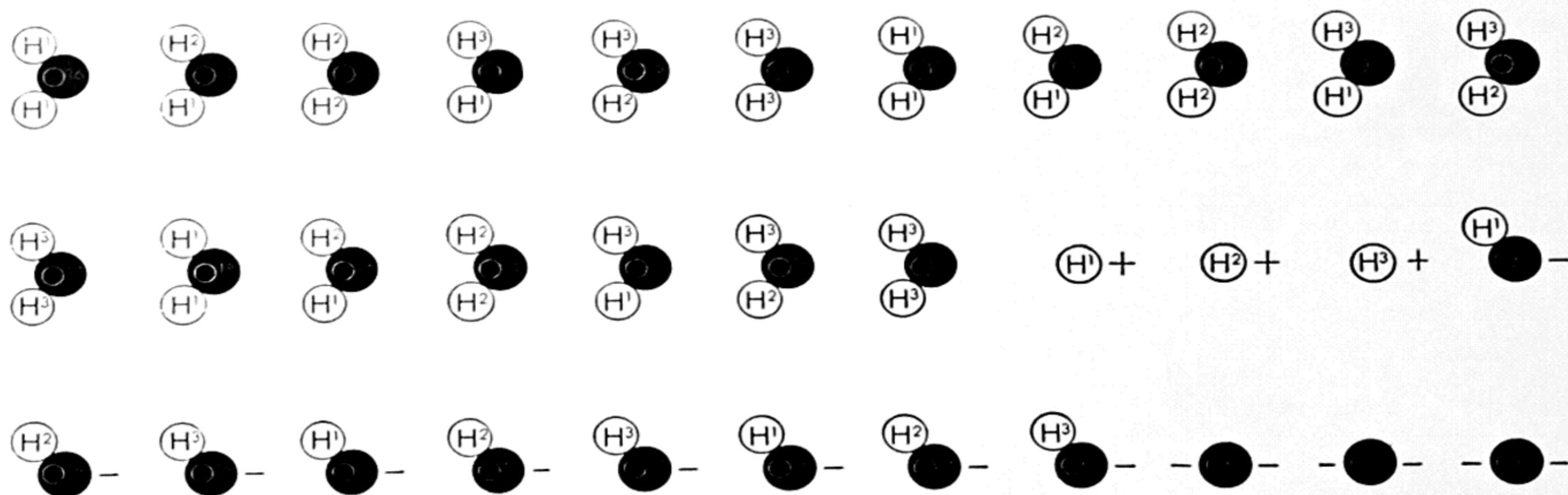


is exhibited on the macroscopic scale in the form of snowflakes. Each molecule is surrounded by four nearest neighbors, so that the group has one molecule at the center and the other four at the corners of a tetrahedron. The molecules and groups of molecules are joined together by hydrogen bonds.

The forces of attraction between the molecules in ice or water produce a strong inward pressure. As we shall see, this accounts for some of water's peculiar properties. In the form of ice, its open structure resembles a bridge arch under heavy downward stress. When the temperature of the ice rises to zero centigrade, the thermal agitation of the molecules is sufficient to cause the ice structure to collapse, and the water becomes fluid. It is well known that the application of pressure from outside will make ice melt at a lower temperature; evidently this reinforces the internal pressure within the ice and assists its collapse. Contrariwise, we can assume that if the internal pressure is reduced in some way, the melting point of ice will rise. Calculations indicate that if this pressure were entirely eliminated, ice would not melt until its temperature reached 15 degrees or more centigrade (59 degrees Fahrenheit).

According to X-ray determinations, the average distance between the center of one oxygen atom and the center of the next in the ice crystal is 2.72 Angstrom units (an Angstrom being one hundred-millionth of a centimeter). When ice melts to liquid water, the hydrogen bonds are stretched and the molecules move farther apart: the distance between oxygens is increased to about 2.9 Angstroms on the average. This stretching would open the structure further and make water less dense were it not for the fact that in the fluid the molecules crowd together in more compact groups. Each molecule is now surrounded by five or more neighbors instead of only four.

The chaotic disorder in which water molecules exist in the liquid state is difficult to picture. Their arrangement shifts continually. The angle between the two hydrogen atoms in the water molecule no longer remains fixed near a right angle but becomes variable, so that the molecule is flexible. Each oxygen atom now attracts by electrical forces not two extra hydrogen atoms as in ice, but three or more. Thus we may find an oxygen atom surrounded by five or six hydrogens and a hydrogen atom surrounded by as many as three oxygens. In the closely knit, flexible structure the hydrogens constantly shift their posi-



WATER IS NOT H₂O but a mixture of 33 different substances. Eighteen of these are combinations of three isotopes of hydrogen and three of oxygen. The three hydrogen isotopes are ordinary hy-

drogen (H¹), deuterium (H²) and tritium (H³). The three oxygen isotopes are ordinary oxygen (O¹⁶), oxygen 17 and oxygen 18. The remaining substances are various ions (*plus or minus signs*).

tions and displace one another. Each such displacement is propagated in a chain or zipper fashion throughout the liquid. This has consequences which affect the viscosity, dielectric constant and electrical conductivity of water.

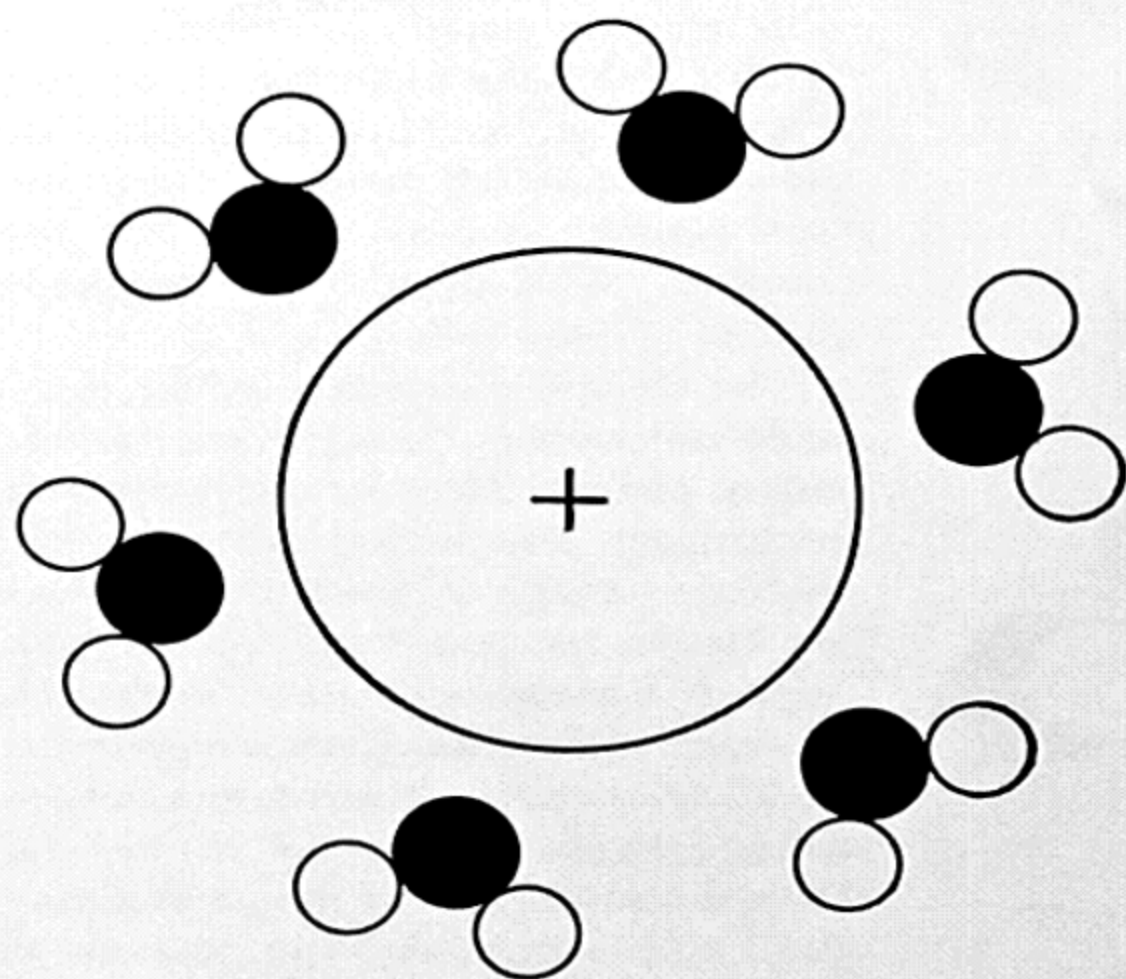
Water's Properties

In an ordinary unassociated liquid such as benzene the molecules flow by sliding around one another. In water the motion is rolling rather than sliding. Since the molecules are connected by hydrogen bonds, at least one bond must be broken before any flow can occur. From the fact that the molecules are bonded together, it might be expected theoretically that the viscosity of water

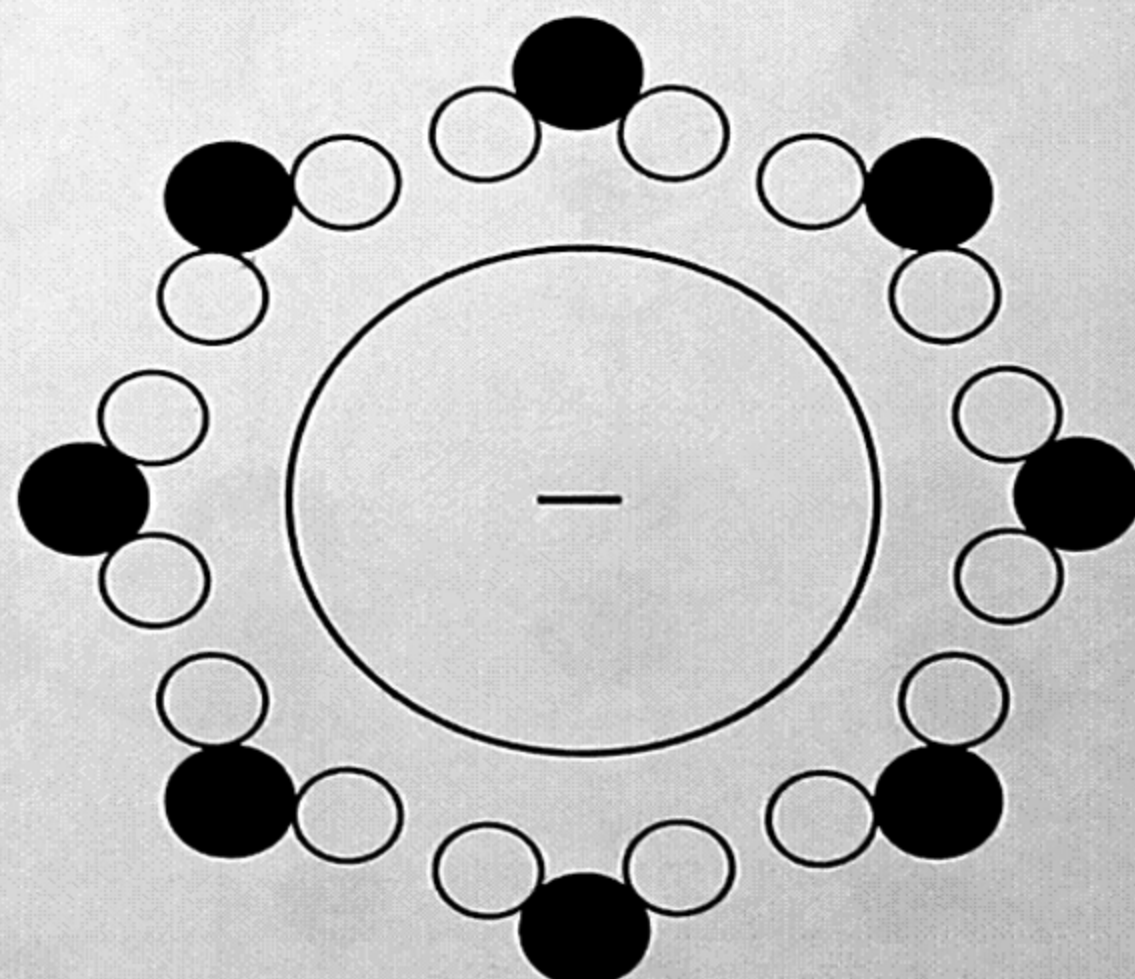
should be comparatively high. However, each hydrogen bond in water is shared on the average between two other molecules, and one of these weakened bonds is easily broken. The greater viscosity of ice is due to the fact that each hydrogen is bonded to only a single oxygen atom from another molecule, and this firmer bond must be broken before any movement can occur.

The dielectric constant of a liquid is a measure of its capacity to neutralize the attraction between electrical charges. For example, when sodium chloride dissolves in water, the positively charged sodium and negatively charged chlorine ions are separated. They are kept apart because water has a high dielectric constant—the highest of any common liquid.

It reduces the force of attraction between the oppositely charged ions in solution to not much more than 1 per cent of the original value. The reason for water's strong neutralizing action lies in the arrangement of its molecules. In an aggregation of water molecules a hydrogen atom does not share its electron equally with the oxygen atom to which it is attached: the electron is closer to the oxygen atom than to the hydrogen. As a result the hydrogen atoms are positively charged and the oxygen negatively charged. Now when a substance is dissolved, separating into ions, the oxygen atoms are attracted to the positive ions and the hydrogens to the negative ones. Consequently water molecules surrounding a positive



IONS (circles labeled with plus and minus signs) in water are kept apart because they polarize the water molecules around them. Because the oxygen atom of the water molecule has more negative



charge than the hydrogen atoms, it is attracted to a positive ion (*left*). Because the hydrogen atoms have more positive charge than the oxygen atom, they are attracted to a negative ion (*right*).

ion are oriented with their oxygens next to the ion, and molecules around a negative ion turn their hydrogens toward the ion. Thus the water molecules act as cages which separate and neutralize the ions. This explains why water is so effective a solvent for electrolytes (substances which dissociate into ions) such as sodium chloride.

Water is generally supposed to be a good conductor of electricity. Every lineman knows the danger of handling high-voltage electrical lines when standing on a moist surface. Actually the conductivity is due to impurities dissolved in the water. Water is such a good solvent for electrolytes, including carbon dioxide from the atmosphere, that any moist surface may be assumed to be a good conductor. But pure water (which is difficult to keep pure—it must be kept out of contact with air and in a vessel of an inert material such as quartz) is a very good insulator indeed. The reason is that while the hydrogen and oxygen atoms in a water molecule are in a sense charged, or ionized, they cannot move about separately because they are attached to each other, and hence cannot carry an electric current.

One of the anomalous properties of liquid water is its high specific heat, or heat-holding capacity. The specific heat of a substance is the quantity of heat required to raise the temperature of one gram one degree centigrade. The specific heat of liquid water is more than twice as great as that of ice. The explanation is that the liquid's ionized oxygen and hydrogen atoms, though held together, behave like free ions in their capacity to vibrate in response to heat. Thus they can absorb as much energy as if the ions were really free.

The strong bonding of water molecules accounts for the fact that water has unusually high melting and boiling points. It also explains why it is so difficult to vaporize ice. To do this we must break all the hydrogen bonds holding the molecules together. Calculations indicate that the total energy of the hydrogen bonds in one mole of water (18 grams) is equivalent to 6,000 calories.

Hydrates

For more than 60 years physical chemists have studied water largely in terms of solutions of electrolytes. This

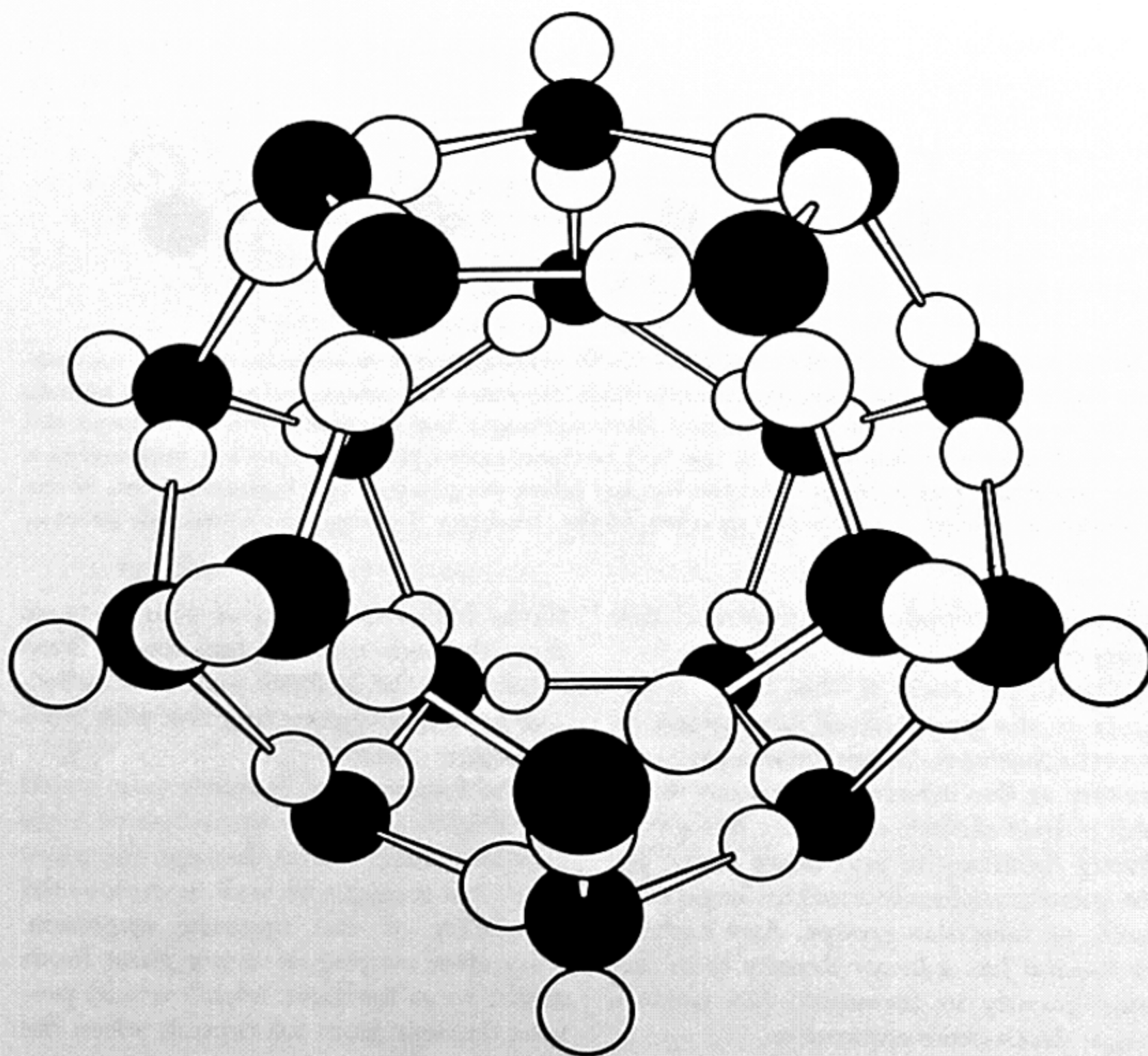
study has produced considerable information about electrolytes and ions, but not a great deal about the properties of water itself. Strangely enough, in recent years we have learned much more about water by examining its behavior with substances which for all practical purposes are insoluble in water!

This behavior was called to the attention of chemists in a dramatic fashion by certain surprising natural phenomena. One was the fact that corn sometimes showed frost effects when the temperature was 40 degrees F., well above freezing. Another was the discovery that pipelines carrying natural gas often became clogged with a slushy "snow," containing water, at temperatures as high as 68 degrees F. The plain indication was that these freeze-ups were due to the water. But this raised some startling and interesting questions. What made water freeze at these high temperatures? How could water combine, or become "bound," with substances which were all but insoluble in it? The mystery was not lessened when it was discovered that even the noble gases such as argon and krypton, which refuse all chemical reactions, could join with water to form a quasi compound.

Let us look at these questions in the light of what we have learned about water's structure and properties. Ten years ago in Illinois we began a study of the water-solubility of certain hydrocarbons. Methane gas will serve as an example. The methane molecule does not form ions in water, nor does it accept the hydrogen bonds. There is very little attraction between it and the water molecule. It is, however, slightly soluble in water, and the dissolving methane molecules form compounds with water—"hydrates"—in which several water molecules are joined to one of methane.

The reaction liberates 10 times as much heat as when methane dissolves in hexane, although it is much more soluble in hexane than in water. This fact becomes even more surprising on close examination. The methane molecule occupies more than twice the volume of a water molecule. To form this relatively large cavity for itself on dissolving, a great deal of energy would be required: it should be somewhat greater than the heat of vaporization of water—say 10,000 calories per mole. How could so much energy be provided? The forces of attraction between methane and water are apparently too slight to supply any appreciable part of such an amount.

There is an alternative possibility. The presence of the methane may drastically change the water structure itself.



HYDRATE is formed when a foreign molecule in water is electrically neutral and just the right size for the water molecules to collect around it in crystalline cage. This cage can then grow to a much larger crystal. It is part of a repeating unit of 136 molecules.

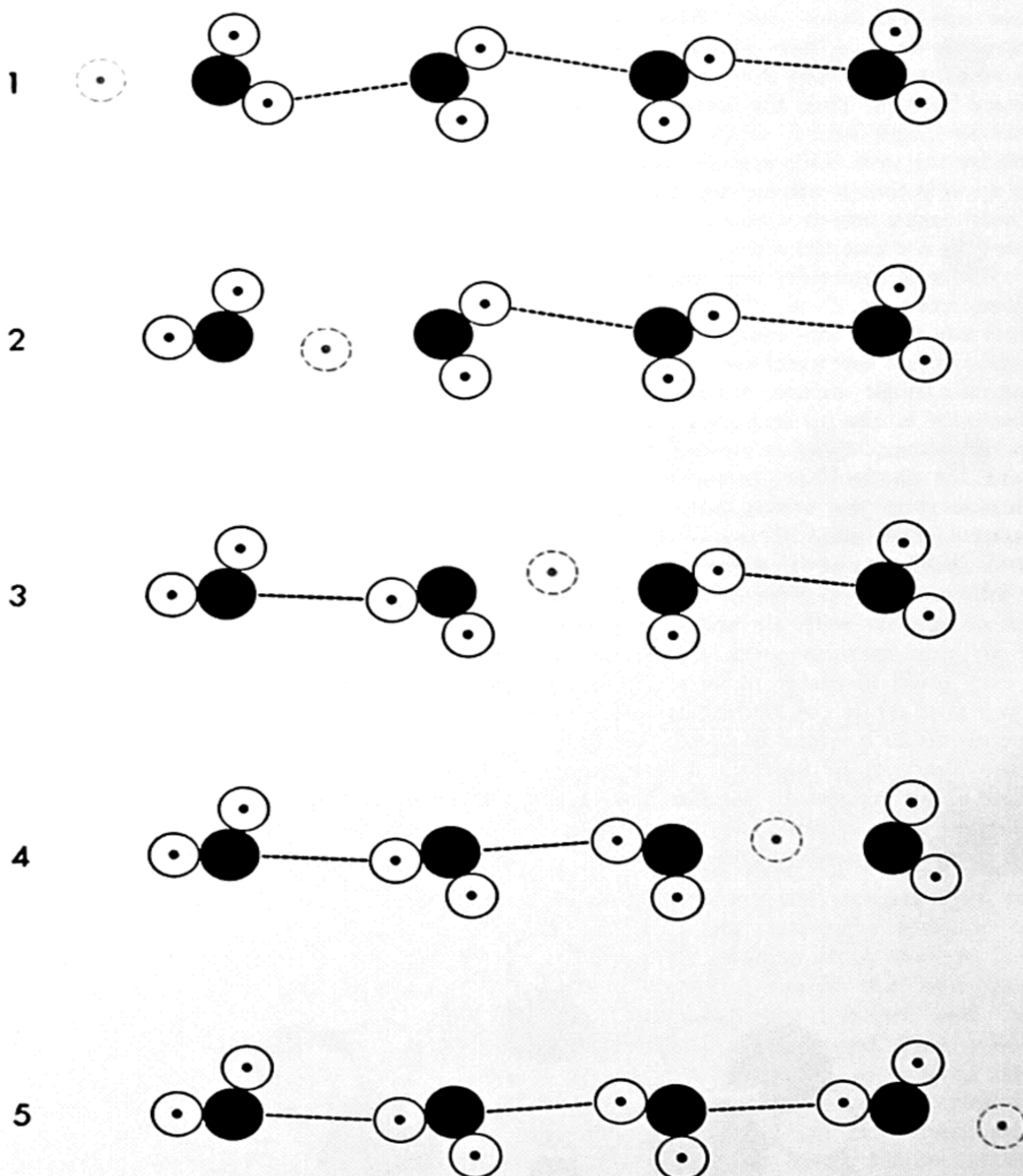
Let us suppose that the dissolved methane molecule is surrounded by an envelope of 10 or 20 water molecules. The formation of such a structure would account for the heat liberated. In the space occupied by the methane molecule the attractive force on the water molecules, and hence the inward pressure, would disappear. Under these conditions, as we have seen, water will freeze at a higher temperature. Thus the molecules at the interface between the methane and water molecules may crystallize into "ice." The frozen hydrates may accumulate and separate out of the solution.

This hypothesis is known as the "ice-berg" theory. It is supported by the fact that practically all the nonelectrolytic substances tested have been found to form solid crystalline hydrates. In contrast, electrolytes show little tendency to form them.

All this leads to an entirely new concept of solubility. Chemists have long supposed that solubility always involves attractive forces. But it now appears that the dissolution of a nonelectrolyte is due not to an attraction between the substance and water but to a lack of attraction. The nonionic substances combine with water because they remove internal pressure and thereby permit formation of a crystalline compound.

In order to understand the formation of these hydrates, it is necessary to consider their molecular structure in detail. They tend to fall into groups according to the number of water molecules they contain.

The ground work for the study of the hydrates structure was laid by M. von Stackelberg in Germany 10 years ago. He showed by X-ray studies that this structure was cubic, in contrast to the hexagonal structure of ice. W. F. Clausen of our laboratory recently attacked the problem of building such cubic structures, each containing a gas molecule, into a repeating lattice. It turns out that there are two possible cubic lattices, one of which was proposed and worked out by Linus Pauling of the California Institute of Technology. This has a spacing of 12 Angstroms between molecules while the other has 17 Angstroms. The smaller lattice contains 46 water molecules and the larger one 136. The holes for gas molecules in the smaller lattice have 12 or 14 walls, while those in the larger one have 12 or 16 sides. These holes are of different sizes and make possible a bewildering array of hydrates. The different sized holes can be filled only with different sized molecules, and not all the holes in a lattice need be filled. The model explains the actual



RAPID MIGRATION OF HYDROGEN IONS through water is explained by the assumption that the ions do not actually travel through the water but are passed from one molecule to the next by a process of exchange. Here hydrogen ion is represented by colored dot surrounded by a broken circle. In the first horizontal row the hydrogen ion approaches a water molecule. In the second row the ion has taken the place of one hydrogen atom of the molecule, expelling the atom as a new ion. In the third row the new ion repeats this process.

composition of hydrates with remarkable precision.

The importance of this type of hydrate to the processes of life cannot be overemphasized. These processes occur mainly at the interfaces between water and protein molecules. Water has a very strong tendency to crystallize there, for the protein molecule contains large non-ionic, or nonpolar groups. Any hydrate so formed has a lower density than ice; consequently its formation can cause a large, destructive expansion.

The freezing of corn at a temperature of 40 degrees F. becomes understandable in terms of the formation of a hydrate. Winter wheat, on the other hand,

forms hydrates slowly as temperatures drop through the fall, and under these conditions the hydrate acts as an effective antifreeze protecting the cells from damage.

The frozen food industry uses rapid freezing to avoid the formation of large crystals which would damage the plant cells. But it might be well to explore the possibility of the opposite approach. Very slow cooling of living plant foods might form hydrates which would prevent damage from ice crystals when the plant was frozen.

Let us return now to see how the structure of water may be modified when an electrolyte, say a salt, goes into solu-

tion. The only direct physical clue we have lies in the behavior of the salt ions in conducting an electric current. The rate of motion of the ions will depend in part on the resistance they encounter in the liquid, and this in turn will depend on the size of the moving particles. If water molecules are attached to an ion firmly enough to move along with it, they will of course increase the apparent size of the ion. Studies of the mobilities of various ions show that positive ions smaller than potassium carry such a cage of water molecules with them. The positively charged ion attracts the oxygens of two water molecules quite strongly, and if the volume of the ion plus its two water molecules is not greater than that of a methane molecule, a cage of hydrogen-bonded water molecules will form around this group as a nucleus. Positive ions larger than

potassium fail to pick up such a cage. The same is true of most, but not all, negative ions.

The positively charged hydrogen ion and the negatively charged hydroxyl ion (OH) are surrounded by cages, and yet they show the highest mobility in carrying a current. We must conclude that they manage in some way to escape from the cage. Actually the mechanism is not hard to picture: they continually form new cages as they travel by a process called proton transfer. Under the influence of an electric field a hydrogen ion may jump from one water molecule to the next. When this has occurred, the hydrogen on the farther side of the water molecule takes up its part in the race like a relay runner and jumps to the next water molecule. Thus, a succession of protons, each doing its bit, carries the current. The motion is rapid, because

each proton moves only a short step. Transfer of the proton also explains the conduction of electricity by hydroxyl ions. When a proton jumps toward the right, say, and joins a hydroxyl ion, it leaves a hydroxyl ion on its left. The effect is the same as if a hydroxyl itself moved to the left.

Water, then, is not simple H_2O but a unique and complicated material with distinct and varied chemical properties. It has a definite though changing physical structure which depends on the orientation of its molecules with respect to one another, and to the molecules of dissolved substances. Since the behavior of all living nature and much of the inanimate world is inseparably linked to the peculiar characteristics of this liquid, the study of water substance can tell us a great deal about fundamental aspects of the world in which we live.

The Authors

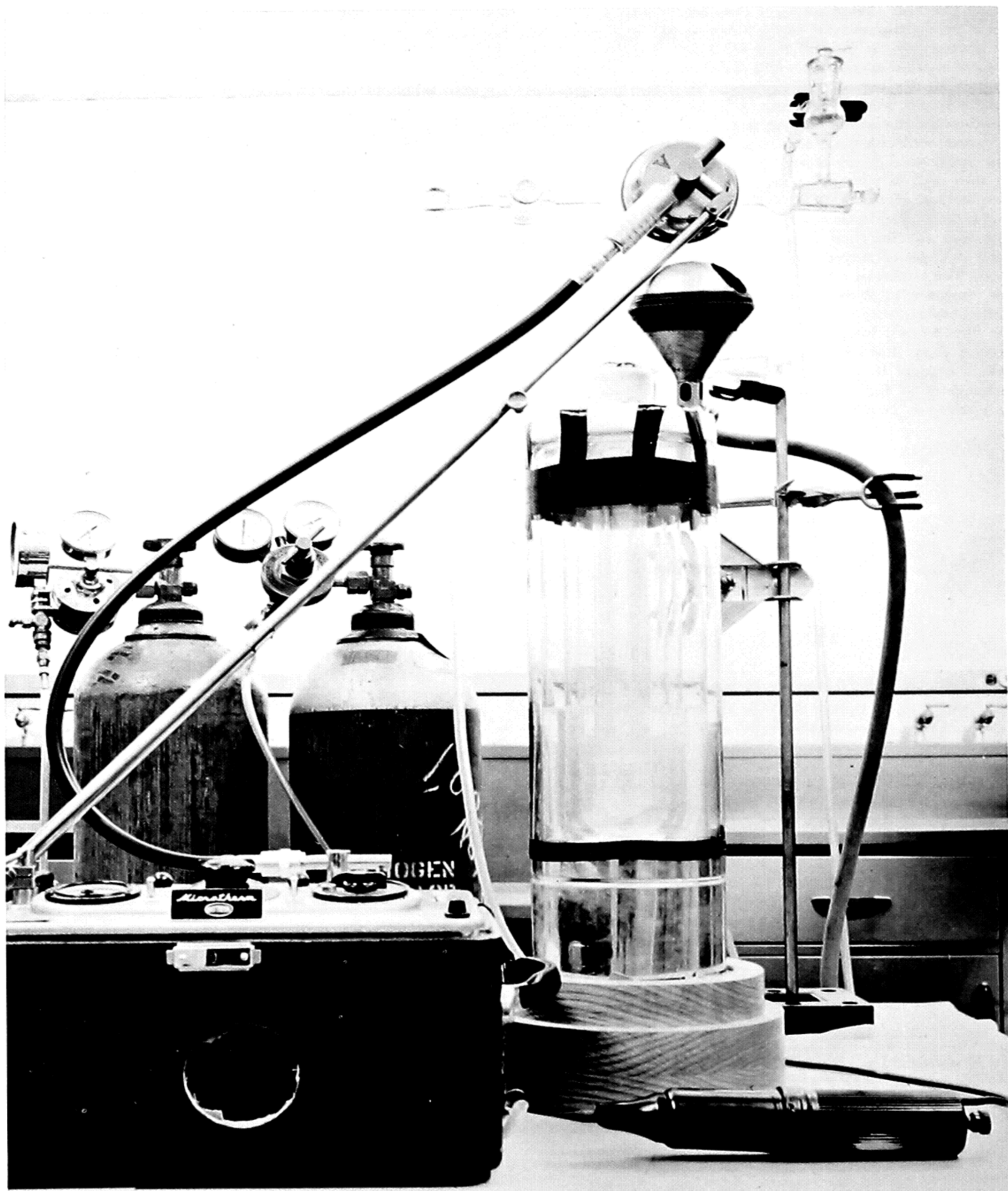
ARTHUR M. BUSWELL and WORTH H. RODEBUSH are chemists who have been doing research on water together since the 1930s. Buswell, now research professor at the University of Florida, was chief of the Illinois State Water Survey and research professor of chemistry at the University of Illinois until September, 1955. He graduated from the University of Minnesota and took his Ph.D. at Columbia University in 1917. At Illinois, from 1920 to 1955, he developed methods of water and waste treatment. Rodebush joined the University of Illinois in 1921 as an asso-

ciate professor of physical chemistry, becoming professor of chemistry in 1924. With Wendell M. Latimer he invented the concept of the hydrogen bond to explain in part the properties of water. The work of Buswell and Rodebush has been supported by grants-in-aid from the National Institutes of Health.

Bibliography

WATER ISN'T H_2O . A. M. Buswell in *Journal of the American Water Works Association*, Vol. 30, No. 9, pages 1433-1441; September, 1938.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.



DEWAR FLASK in which radicals are frozen is the vertical glass structure at the right. The box in the left foreground is a high-

frequency generator. Behind it are tanks of gaseous nitrogen and helium. The cylinder lying in front of the flask is a spark coil.

FROZEN FREE RADICALS

by Charles M. Herzfeld and Arnold M. Bass

Molecular fragments, which normally have a fleeting existence during chemical reactions, can now be preserved by rapid cooling. The study of their properties is an exciting field of research.

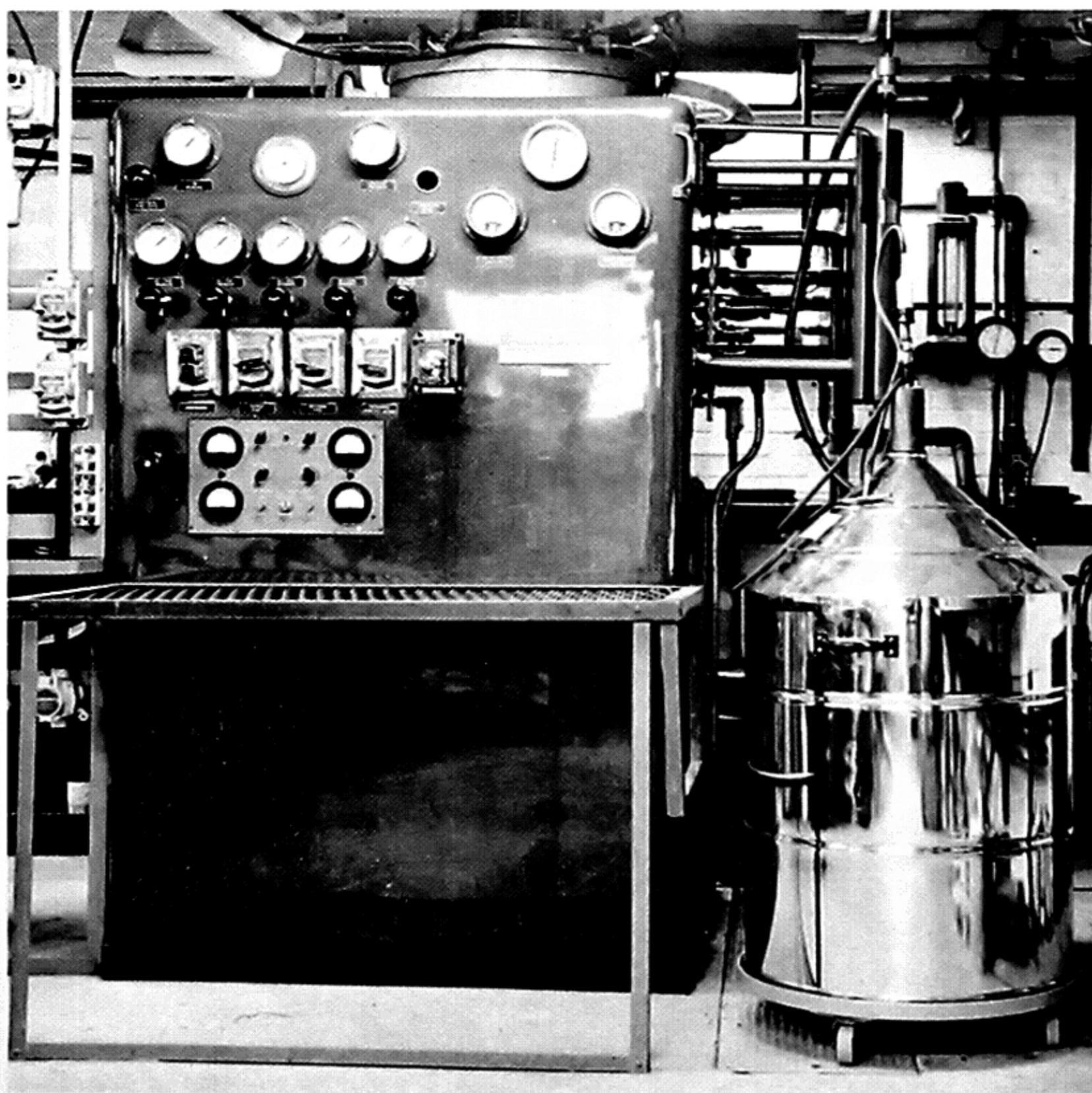
Almost everywhere we look in the universe we can see the dance of free radicals. They are present in an ordinary flame, in an electric arc, in the atmosphere, in the stars, even in the cold interstellar dust. Free radicals are fragments of matter which for the most part have only a fleeting existence. If we understood their behavior fully, we would have a master key to understanding what makes the chemistry of the universe go. One of the most exciting physicochemical developments of recent years is the discovery of ways to halt the dance of the free radicals—to freeze them in their tracks so that we can examine them at our leisure. What the investigators have seen already is so startling and so stirring to the imagination that it is no exaggeration to say it has opened a new field of science.

What is a free radical? We can illustrate what it is and what role it plays by considering an atom of chlorine. The atom may be freed by sunlight breaking up a molecule of chlorine gas (Cl_2). In the presence of hydrogen, the chlorine atom attacks a hydrogen molecule (H_2), forming the comparatively stable product HCl and releasing a hydrogen atom. Now it is the hydrogen atom's turn to act as a free radical: it attacks a chlorine molecule, again producing HCl and freeing a chlorine atom to continue the chain reaction. This in simple outline is the role of free radicals—which may be either single atoms or larger fragments of molecules. The highly reactive fragments start and maintain a chain reaction among comparatively inactive substances. A relatively small number of free radicals (as few as one per thousand molecules) can keep a chain reaction running. Because each step takes only a small fraction of a second, they turn out products at a great rate.

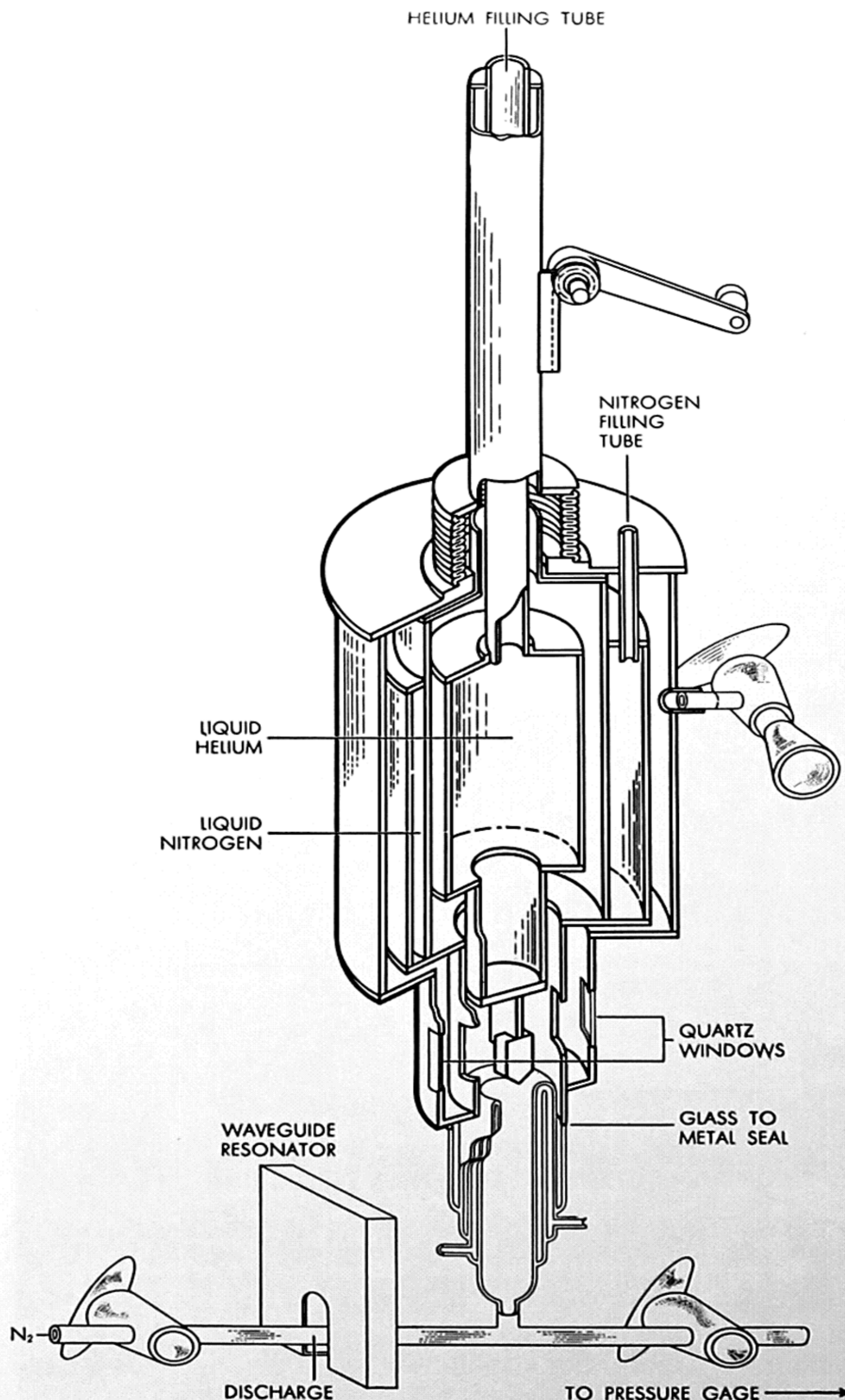
Free radicals almost invariably contain an odd number of electrons, which in part accounts for their activity. (Hydrogen has one electron; chlorine has 17; the hydroxyl radical— OH —has nine; the methyl radical— CH_3 —has nine, and so on.)

Not all free radicals are short-lived. There are species which have lifetimes of

many days under suitable conditions—e.g., when they cannot react with air. The known long-lived types are mainly hydrocarbons, usually consisting of 30 or more atoms, which can be kept by dissolving them in benzene or similar inert solvents [see illustration on page 519]. Such radicals were first isolated half a century ago by Moses Gomberg of the



LIQUID HELIUM for the experiments at the National Bureau of Standards was made by this helium cryostat. The vacuum storage tank at right holds 20 gallons of liquid helium.



FREEZING APPARATUS of the Bureau of Standards is a modified version of one developed at the Johns Hopkins University Applied Physics Laboratory. Nitrogen (N_2) is admitted at lower left and is dissociated by the high-frequency field of the waveguide resonator. Radicals collect on a copper wedge inside the quartz windows. The wedge is connected to a helium reservoir by a copper rod. Liquid nitrogen in outer container insulates the helium.

University of Michigan [see "Free Radicals," by Paul D. Bartlett; *SCIENTIFIC AMERICAN*, December, 1953].

But it is the short-lived radicals, whose lifetimes are usually measured in thousandths of a second, that are most abundant and most interesting. They are generally simple in structure and consist of only a few atoms: the hydroxyl and methyl radicals are typical. Free radicals are generated wherever molecules are broken up by some energetic process: a chemical reaction, heat, an electrical discharge, the attack of ultraviolet light and cosmic rays on molecules in our upper atmosphere. The dust of interstellar space is probably made up of free radicals frozen in the radical state by the extreme interstellar cold.

One of the first scientists to study frozen free radicals was the Norwegian physicist L. Vegard. He believed that certain unexplained wavelengths of the light of the aurora borealis were emitted by particles of solid nitrogen under bombardment by cosmic rays in our upper atmosphere. To investigate the matter he froze some nitrogen at the temperatures of liquid hydrogen and liquid helium, and then bombarded it with very fast electrons and with alpha particles. He did not get the wavelengths he was looking for, but he did produce interesting emissions of light, and he continued his experiments with frozen oxygen, hydrogen and many other materials, including mixtures of various gases. This work, done between 1925 and 1935, unfortunately did not receive sufficient attention. It is now clear that Vegard's experiments revealed some of the properties of free radicals.

In 1948 F. O. Rice and his students at the Catholic University of America hit upon a technique for trapping free radicals and began a series of remarkable investigations. Rice generated the radicals by passing gases through a hot tube (among other ways) and then trapped the products on a cold wall [see "The Chemistry of Jupiter," by Francis Owen Rice; *SCIENTIFIC AMERICAN*, June, 1956]. He obtained and studied some very unusual free radicals, including sulfur (S_2) and nitrogen hydride (NH).

A number of other laboratories are now investigating frozen free radicals by various techniques. This article is principally a report on some recent work at the National Bureau of Standards, initiated by Herbert P. Broida and John R. Pellam in 1954.

They produced free radicals by flowing gases through a high-frequency elec-

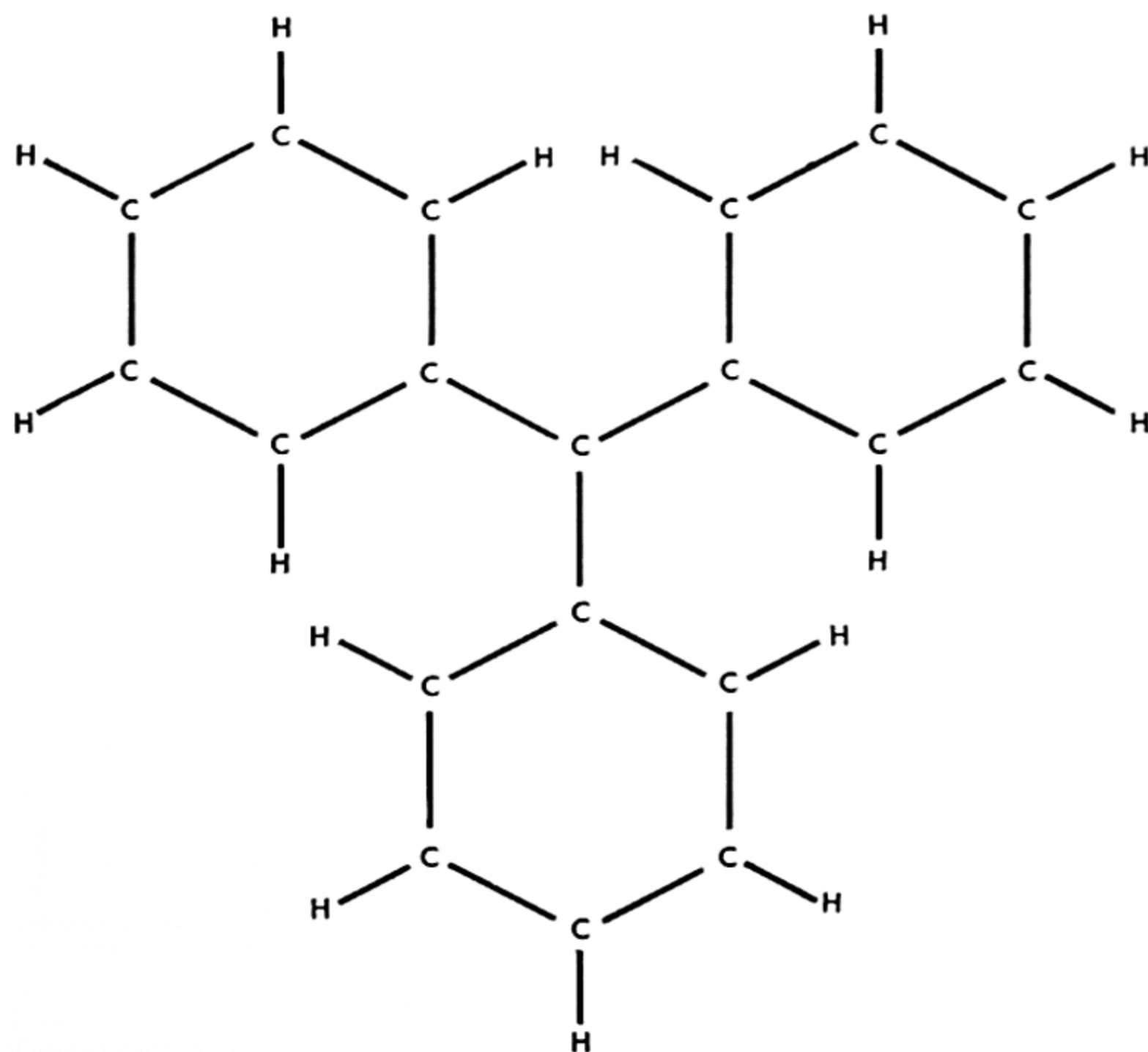
tric discharge, powered by a 2,450-megacycle medical diathermy unit. This convenient system not only produced considerable numbers of free radicals but also raised them to highly excited states. The products were then rapidly pumped to a surface cooled by liquid helium to about four degrees above absolute zero, where they immediately froze into a solid.

Broida and Pellam first tried a few rapid experiments to test their technique. In one exciting day they obtained a series of spectacular breakdown products from discharges through water vapor, oxygen, nitrogen and hydrogen. Broida and his collaborators proceeded to study the properties of these strange new materials. The experiments of that first day provided grist for years of fascinating investigations.

The most dramatic of the first experiments was the one on nitrogen. When the material produced by the discharge breakdown of nitrogen gas began to freeze on the cold surface, it began almost immediately to emit a bright green glow, so intense that it was visible in a well-lighted room. As more material was deposited on the surface, it gave forth brilliant blue flashes. After the flow was stopped, a blue-green afterglow persisted in the cold solid for several minutes. If the solid was then warmed suddenly (to above 25 degrees absolute), it emitted a flash of blue light, which looked like a flame "burning" through the solid. Upon later recooling to the temperature of liquid helium, a blue-green afterglow appeared again, but now much weaker.

This totally unexpected display naturally demanded immediate analysis with the spectroscope, to identify the substances responsible for it. The spectra showed five closely spaced lines of intense emission in the blue-green region, three broad bands in the yellow region and many weaker bands distributed over the whole visible spectrum (a pattern just like the one Vegard had found in his nitrogen experiments nearly 30 years earlier). There was little doubt that most of the green light came from isolated, excited nitrogen atoms, and the presence of such atoms has recently been confirmed by studies of the magnetic properties of the material by C. K. Jen and S. N. Foner at the Applied Physics Laboratory of the Johns Hopkins University. Other features in the spectra were identified as originating from NO, NH and excited nitrogen molecules.

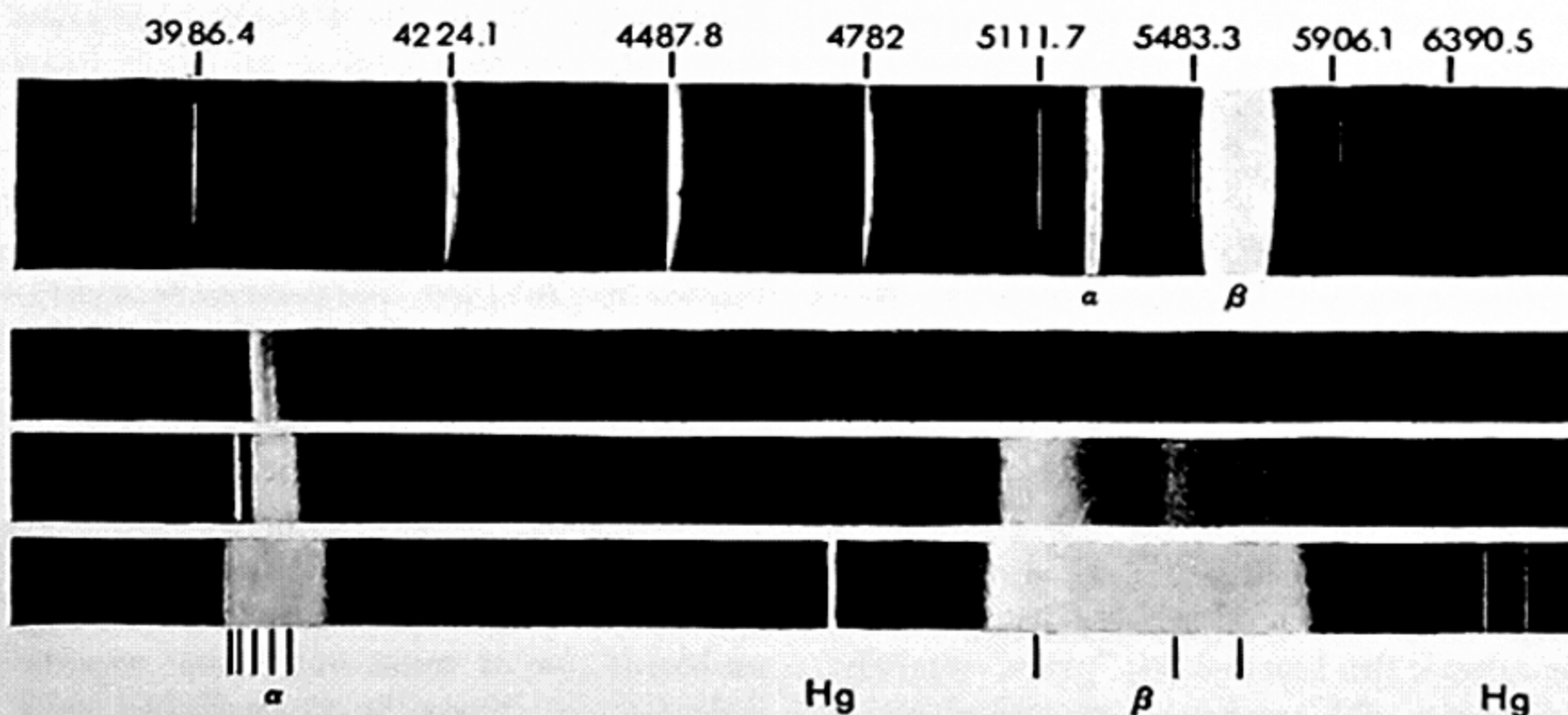
Further information has been obtained by heat measurements. For this



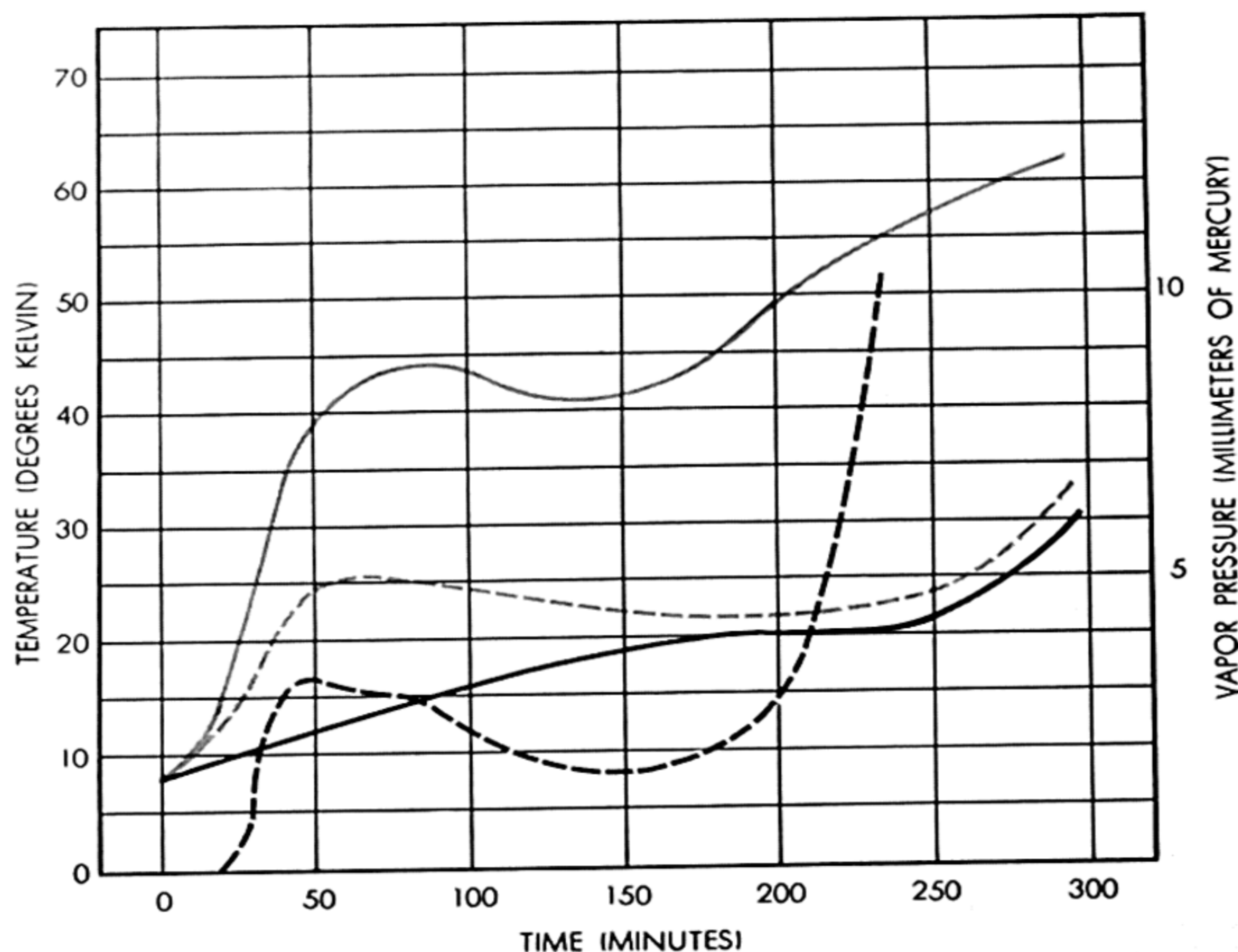
LONG-LIVED RADICAL is typically a hydrocarbon with perhaps 30 or more atoms. This diagram shows triphenylmethyl, first made by the pioneer investigator Moses Gomberg.



SHORT-LIVED RADICALS contain only a few atoms. At left is the methyl radical; at right, the hydroxyl radical. Their lifetimes are normally only a few thousandths of a second.



SPECTRUM OF GREEN GLOW from solidified nitrogen radicals (*above*) has two especially bright bands, marked "alpha" and "beta." These are magnified (*below*) to show five lines in the alpha band and three in the beta. These lines are emitted by excited nitrogen atoms. Lines marked "Hg" are from mercury spectrum, superimposed to help in measuring wavelengths. Remaining lines in upper spectrum, whose wavelengths in Angstrom units appear at the top, are emitted by nitrogen molecules formed by recombination in the solid.



WARMING OF SOLID NITROGEN from a gas discharge is shown by the solid gray curve. The rapid increase in temperature in the first hour is probably due to heat released by recombining atoms. The broken gray curve and solid black curve show warming behavior for solid molecular nitrogen and an empty chamber, respectively. Broken black curve shows how the pressure of gas in the chamber increases as the solid warms and vaporizes.

purpose the products are trapped in a small, low-temperature calorimeter. As the temperature of the solid is allowed to rise, the frozen radicals gradually gain in mobility, until they are able to diffuse through the solid with sufficient freedom to find partners with which to combine. In so doing they emit considerable heat. From the heat liberated, the number of free radicals originally trapped can be estimated. In the case of nitrogen the proportion of free nitrogen atoms in the condensed solid amounts to at least one fifth of 1 per cent.

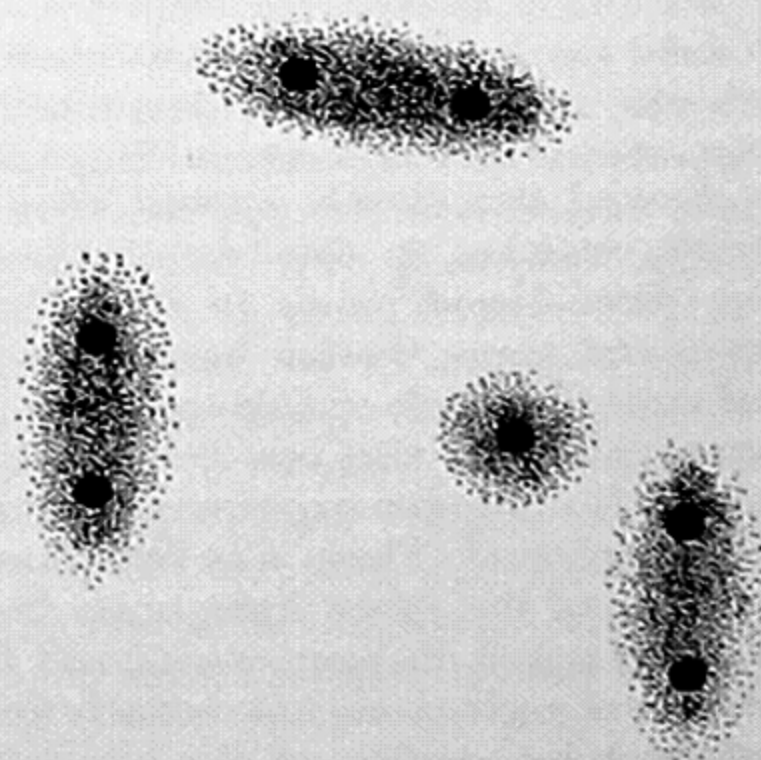
All the experimental evidence indicates that at four degrees absolute free radicals survive for many hours in their isolated state. The reasons for this are not yet precisely clear. The most likely reason is that in the solid material condensed on the cold surface, the free radicals are separated from one another by unreactive molecules and atoms in a rigid framework. When the temperature is raised, the framework "gives" slightly, allowing the radicals to wander far enough to combine with each other. Thus two nitrogen atoms may meet and form the stable nitrogen molecule. In this process they give off just as much energy as is required to tear apart the

two atoms of a nitrogen molecule. About half of the energy released on combination is radiated as light by the newly formed molecule, and the rest is in the form of heat. The color of the light emitted is a measure of the energy required to break up a given molecule.

The bright blue-green glow which is so spectacular a feature of the freezing of nitrogen atoms gives us considerable information about the way the atoms are trapped. The five intense emission lines and the measured lifetime of the after-glow are interpreted to mean that each nitrogen atom is trapped near a nitrogen molecule in a particular geometrical relation to it [see diagram at right]. This interpretation is based on calculations of the disturbance of the motion of the electrons in the atom by the electrons in the molecule. The calculations, and magnetic experiments by Jen and Foner, indicate that the disturbed nitrogen atom is not bound to the nearby molecule, or at most only very weakly linked to it. There is a possibility that "super-nitrogen" molecules, consisting of three nitrogen atoms tightly bound together (N_3), may be formed by the free radicals, but so far we have no evidence of this. In any event, such a mole-

cule could not emit the five blue-green spectral lines.

While the nitrogen phenomena are the most colorful, the experiments with oxygen produced results which in some ways are no less interesting. When molecules of oxygen gas are broken up by an electrical discharge, the material trapped on the cold wall is very complex and variable. Because it emits no light, it has been investigated through its absorption of light. Under some conditions it forms a clear, glassy deposit which absorbs light in a rather complex way—some 25 separate bands or lines of absorption distributed over the whole visual spectrum and extending into the ultraviolet and the infrared. When warmed to about 20 degrees absolute, the deposit changes to a violet-colored solid. This has been identified as a mixture of oxygen and ozone, with a fairly high concentration of ozone. Further warming of the violet substance produces additional ozone in large quantities. By this method about 30 per cent by weight of the oxygen that originally flowed through the discharge has been converted into ozone. Conventional ozone-producing equipment seldom yields more than 6 per cent of ozone from oxygen. Since ozone is an important chemical with many uses, the frozen-radical method may have commercial possibilities as a production process.



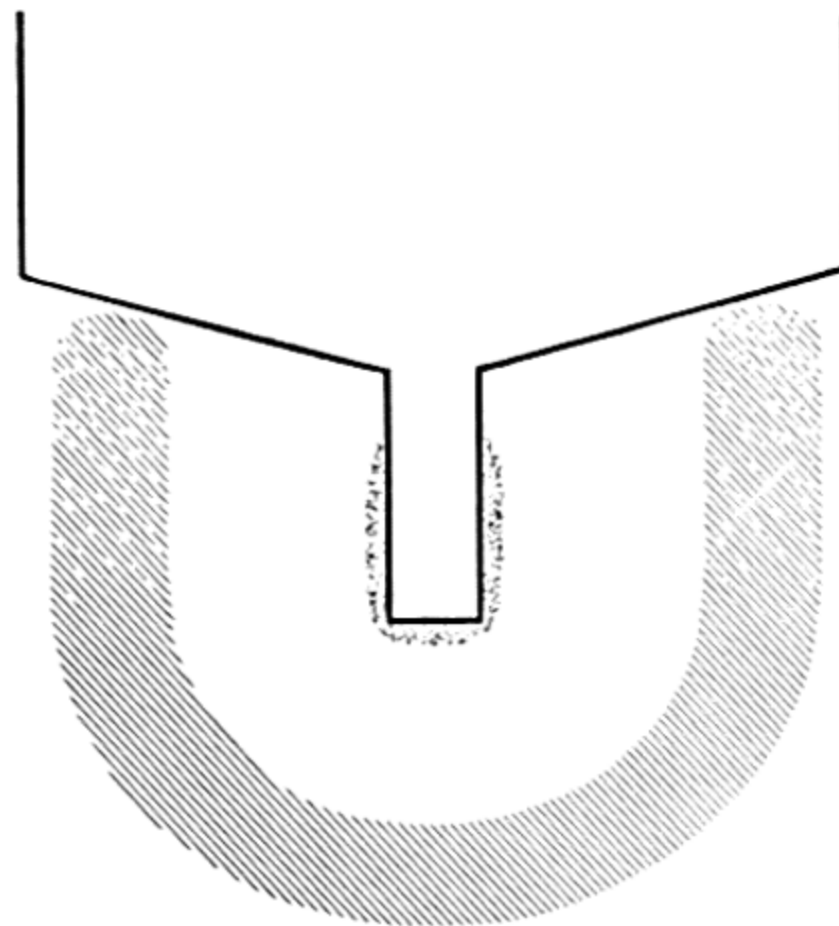
TRAPPED ATOM (single dot) is shown among several nitrogen molecules (double dots). The light emitted by the frozen solid indicates that each atom is located much nearer to one molecule than to the others.

Besides nitrogen and oxygen, the Bureau of Standards laboratory has studied free radicals produced from hydrogen, water vapor and ammonia gas. The products trapped after the breakdown of hydrogen molecules emit no visible light, but the gas surrounding the deposit emits a blue-green glow, quite different from that of the nitrogen radicals. The deposit has been proved to contain free hydrogen atoms, most directly by experiments by Jen and Foner. The breakdown of water yields unusual substances which absorb light in a peculiar way, but they have not yet been identified. The trapped products from ammonia gas emit a blue glow which disappears as soon as the discharge is stopped.

The experiments described in this article have barely scratched the surface of a vast field. By such studies we can expect to obtain a clearer picture of

the roles free radicals play in electric arcs, flames, interstellar dust, comets, stars and galaxies. The method further provides a powerful new tool with which to probe the structure of solids. The spectra of the light emitted by the free radicals trapped in a solid can give information about the arrangement of the atoms and molecules in the solid, about the forces acting on them, about the movements of atoms and about reactions between atoms and molecules.

The efficiency of the method described for producing ozone suggests the possibility of a new industry of very low temperature chemistry. It may eventually become possible to make new chemical compounds having quite unusual characteristics, and to find more efficient methods of producing familiar substances, such as ozone and hydrogen peroxide (H_2O_2), through the control of free radicals.



FROZEN HYDROGEN (stippled area) collects on metal cooled by liquid helium. It is surrounded by glowing gas (hatching).

The Authors

CHARLES M. HERZFELD and ARNOLD M. BASS collaborate on research at the National Bureau of Standards, where Herzfeld is acting assistant chief of the heat and power division. Herzfeld was born in Austria in 1925, came to the U.S. during the war, and in 1945 graduated *cum laude* from Catholic University. He then studied physical chemistry at the University of Chicago, where he obtained his Ph.D. with a thesis on the magnetic properties of uranium compounds. Bass is assistant chief of the free-radicals research section of the Bureau of Standards. A graduate of the College of the City of New York, Bass received his Ph.D. from Duke University in 1949 after World War II service as radar officer aboard a seaplane tender. His main interest has been spectroscopy, particularly in studying the energy states of free radicals in flames for the purpose of temperature measurement.

Bibliography

- THE FORMATION OF THE IMINE RADICAL IN THE ELECTRICAL DISCHARGE. Francis Owen Rice and Melvin Frearno in *Journal of the American Chemical Society*, Vol. 75, No. 3, pages 548-549; February 5, 1953.
- THE IMINE RADICAL. Francis Owen Rice and Melvin Frearno in *Journal of the American Chemical Society*, Vol. 73, No. 12, pages 5,529-5,530; December, 1951.
- INTERPRETATION OF SPECTRA OF ATOMS AND MOLECULES IN SOLID NITROGEN CONDENSED AT 4.2° K. C. M. Herzfeld and H. P. Broida in *The Physical Review*, Vol. 101, No. 2, pages 606-611; January 15, 1956.
- SPECTRA EMITTED FROM SOLID NITROGEN CONDENSED AT 4.2 K. FROM A GAS DISCHARGE. Arnold M. Bass and Herbert P. Broida in *The Physical Review*, Vol. 101, No. 6, pages 1,740-1,747; March 15, 1956.
- THE STRUCTURE OF MATTER. Francis Owen Rice and Edward Teller. John Wiley & Sons, Inc., 1949.
- STRUCTURAL INORGANIC CHEMISTRY. A. F. Wells. Oxford University Press, 1950.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

THE EXCLUSION PRINCIPLE

by George Gamow

It governs both matter and antimatter, explains the internal structure of atoms and nuclei, and enables us to predict the behavior of a confusing assortment of fundamental particles.

One of the most powerful generalizations of modern physics states that each quantum orbit of an atom can contain no more than two electrons. It is sometimes called the Pauli principle, but its author, the theoretical physicist Wolfgang Pauli (who died in Switzerland in December 1958), preferred to call it the exclusion principle. This may have been due to the fact that he did not wish to have it confused with the Pauli effect, a contribution to experimental physics of which he was especially proud.

It is well known that theoretical physicists are quite inept in handling experimental apparatus; in fact, the standing of a theoretical physicist is said to be measurable in terms of his ability to break delicate devices merely by touching them. By this standard Wolfgang Pauli was a very good theoretical physicist; apparatus would fall, break, shatter or burn when he merely walked into a laboratory. The explosion of some elaborate vacuum equipment in James Franck's laboratory at the University of Göttingen has been directly attributed to the Pauli effect: it was later definitely established that the mishap had occurred at the exact time a train carrying Pauli had stopped momentarily in the Göttingen railroad station.

Pauli's exclusion principle, on the other hand, acquired its importance because it helped to clarify the internal structure of the atom. According to Niels Bohr's model of the atom, electrons circling an atomic nucleus may move only along certain "quantized" orbits. (Any variable in nature restricted to a series of discrete values is said to be quantized.) Electrons can emit radiation by jumping from the outer orbits (which represent higher energies) to inner ones. The illustration on page 524

shows the quantum orbits in the hydrogen atom along which that atom's single electron is allowed to move.

Closest to the nucleus is the first quantum orbit, a circular one which corresponds to the lowest energy-level; it is the permanent abode of the electron in its normal state. When the electron is excited to the next higher state of energy, it may reside in the second energy-level, in either the second circular orbit or any of three associated elliptical orbits, all of which have the same energy. A still more energetic electron may reside on the third energy-level, which consists of the third circular orbit and eight elliptical ones, and so on. Successive sets of orbits (or "electron shells," as they are usually known) correspond to ever higher energy-levels and have an ever increasing number of elliptical orbits associated with the principal circular ones.

Orbits and Periodicity

In the atoms of heavier elements (whose nuclei have larger electrical charges) the consecutive electron shells are arranged in the same pattern as those in the hydrogen atom, but the diameters of the shells are somewhat smaller. In these atoms the increased electrical attraction of the proton-packed nucleus is not completely balanced by the increased electrical repulsion of the negatively charged electrons, so the electrons are pulled closer to the nucleus. This presents something of a problem, because heavier nuclei have an increasingly larger number of electrons orbiting around them. In the oxygen atom, for example, there are eight electrons instead of the hydrogen atom's one; in uranium there are 92 electrons. We might then ask: How is the larger number of electrons in the atoms of the

heavier elements accommodated in the smaller quantum orbits?

In terms of classical physics the answer to this question is almost trivial. The most stable state of any mechanical system is the one in which the system cannot lose any more energy by dropping to a still-lower energy level. Thus all the additional electrons in heavier atoms might be expected to drop into the first quantum orbit and play "ring-around-a-rosy," or, more exactly, ring around the nucleus. And because we know that the diameter of that ring becomes smaller in heavier elements, we might predict that it would also become more and more tightly packed with electrons. If this were true, the atoms of hydrogen, oxygen and uranium would look as shown in the illustration at the top of page 525. The fact is this does not happen: regardless of the charge of the nucleus, the over-all size of atoms remains approximately the same.

It was to explain this situation that Pauli first postulated his now-famous principle that each quantum orbit may hold no more than two electrons. The Pauli principle showed further that if both these vacancies are filled, the next electrons must be accommodated in other orbits. When all the orbits in a given shell are filled, the orbits in the next shell (corresponding to a higher energy-level) begin to fill. Thus although the diameters of quantum orbits are smaller in heavier elements, a steadily increasing number of them is filled up. This explains why all atoms are of roughly the same size.

The electron shells of all the species of atoms in the periodic table are filled according to this fixed hierarchy of energy states. The first shell, which represents the lowest available energy-state, is the first to fill. In the helium atom this



WOLFGANG PAULI, formulator of the exclusion principle, received the Nobel prize in physics for it in 1945. A brilliant theoretical physicist, Pauli was also famous for his work in particle

physics (he postulated the existence of the neutrino) and in quantum electrodynamics, the set of theories that describes the behavior of electrons in electromagnetic fields. He died in December, 1958.

shell is completely filled by the two electrons chasing each other around the first quantum orbit. The next element, lithium, has three electrons, one of which, according to the exclusion principle, must be added in the second shell, consisting of one circular and three elliptical orbits. Since these four orbits can hold a total of eight electrons and the inner orbit holds two, both the first and second shells will be filled in the neon atom, which has 10 electrons. The extra electrons in still-heavier elements must be added in a third set of circular and elliptical orbits, and so on. Pauli's exclusion principle thus explains the internal structure of elements in terms of the way in which their consecutive electron shells

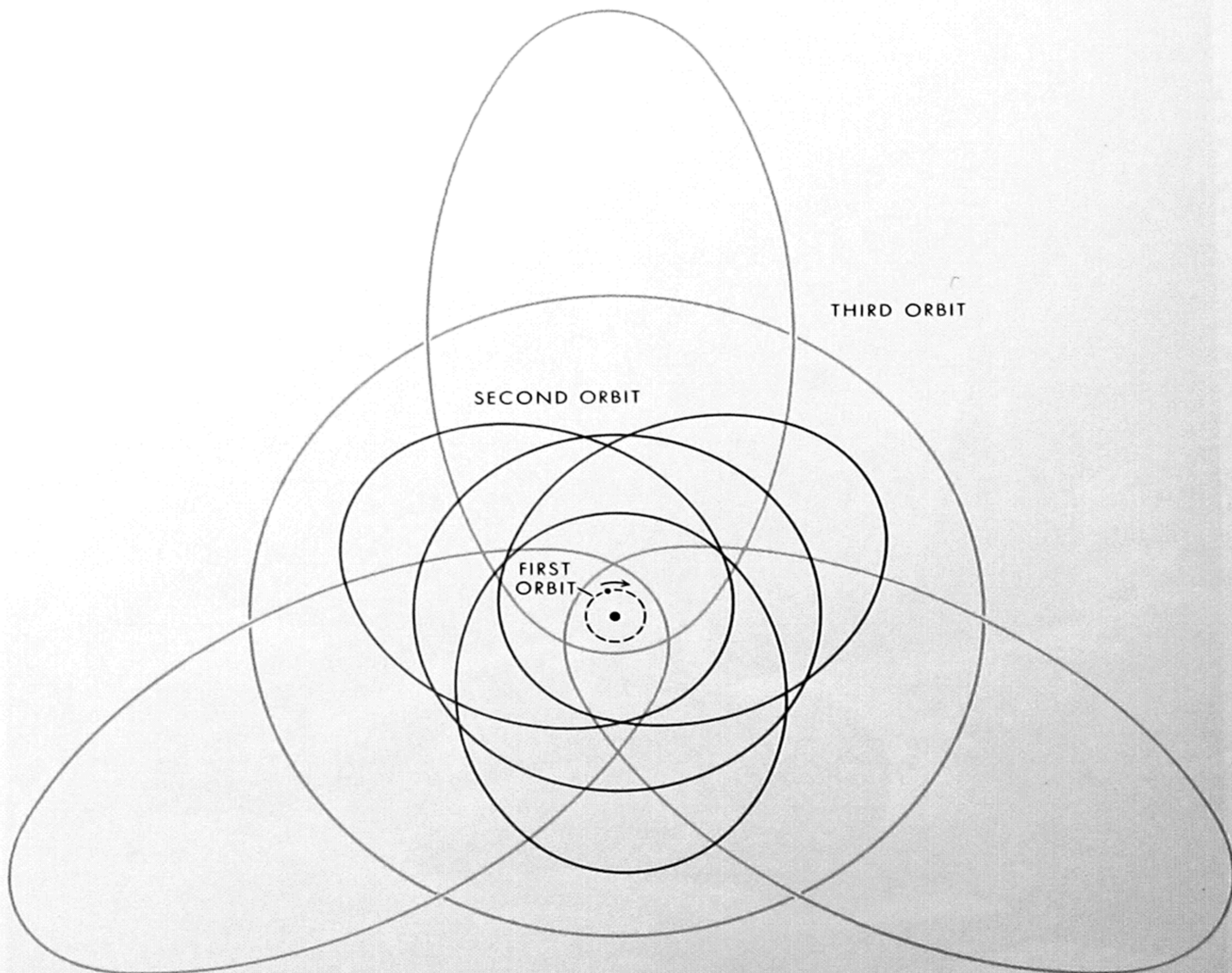
are filled. The principle also underlies the external, or chemical, identity of an atom and the periodicity of chemical properties in the sequence of atomic species in the table of elements. These characteristics are determined by the number of electrons in the outer shells of atoms, which make contact when atoms collide with one another.

Electron Spin

When the Pauli principle was originally formulated, electrons were believed to be nothing more than point charges of negative electricity. It was soon discovered, however, that electrons must also be considered as tiny magnets: they pos-

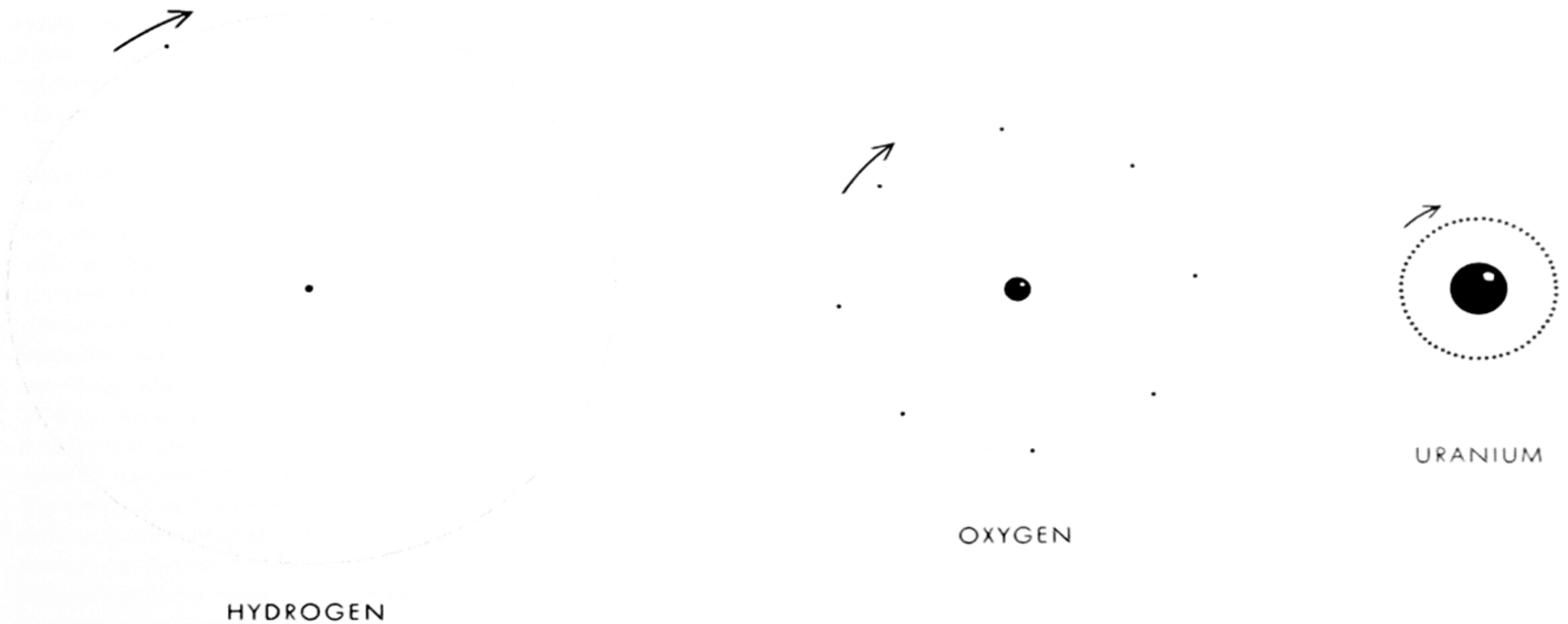
sess a magnetic moment because they spin rapidly as they orbit around a nucleus. Having learned to regard electrons as tiny magnets, we must take into account both the electrical forces which are mainly responsible for their orbital motion and the magnetic forces set up by their spin.

An electron tends to spin in one of two ways: either in the direction in which it travels along its orbit, or in the opposite direction. It was shown that two electrons that follow the same orbit must spin in opposite directions [see illustration at lower right]. This discovery requires us to formulate the Pauli principle in a somewhat different way. Because electrons spinning in op-



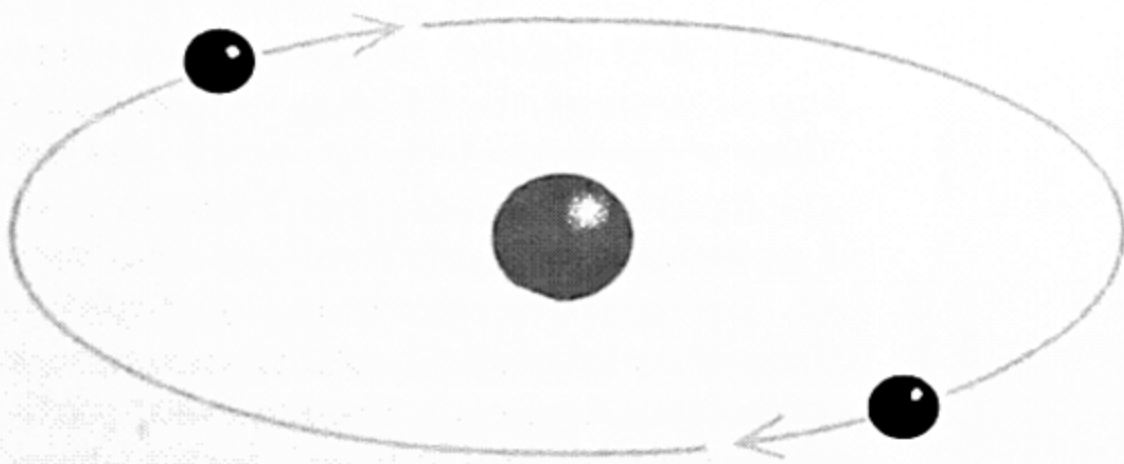
BOHR MODEL OF HYDROGEN ATOM contains consecutive sets of quantum orbits along which a single electron moves around a one-proton nucleus. First circular orbit (*broken line*) represents the lowest energy-level and is the normal "home" of the electron.

The second set of orbits (*color*) represents the next energy level, and the third set (*gray*) a still-higher level. Only the first three sets of orbits are shown; the additional sets have an increasing number of elliptical orbits associated with the principal circular ones.

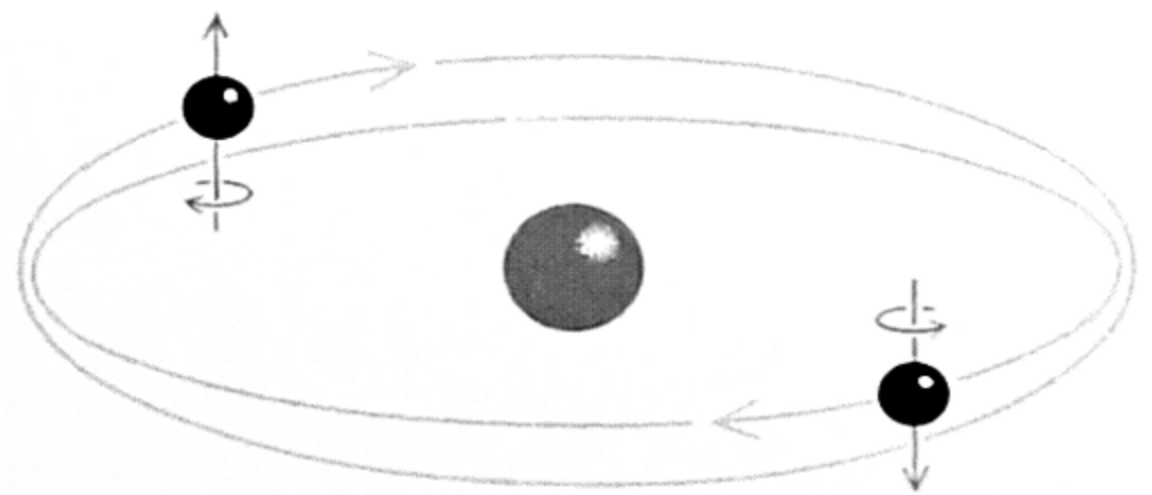


ATOMS WOULD DIFFER IN SIZE if the exclusion principle did not apply. The laws of classical physics would require all of an atom's electrons to occupy the first quantum orbit because it represents the lowest energy-state. The laws also predict that this orbit

would be smaller in heavier atoms, because of increased electrical attraction between nucleus and electrons. Thus the atoms of hydrogen, oxygen and uranium would appear approximately as shown here. Actually all three atoms are approximately equal in size.



ORIGINAL EXCLUSION PRINCIPLE permitted two electrons (*black*) to move around a nucleus (*gray*) in the same orbit. Electron's spin and magnetic moment had not been discovered.



MODIFIED EXCLUSION PRINCIPLE, necessitated by the discovery of electron's magnetic moment, states that the two electrons must spin in opposite directions and move in two separate orbits.

posite directions set up weak magnetic fields which slightly alter each other's orbits, we now say that the two electrons originally permitted to travel in the same orbit actually follow two different (though very similar) orbits. It is therefore more rational to regard the permitted orbits as close pairs split apart by weak magnetic interactions.

Pauli's exclusion principle applies not only to electrons in atoms but also to "free" electrons which have abandoned their atoms to drift freely through matter. We know that the electrical conductivity of a metal is due to the free electrons traveling through the metal's crystal lattice. When a voltage is applied to the metal, these free electrons move preferentially in the direction of the electric force acting on them, and comprise an electric current. We often speak of them as an "electron gas" which permeates the metal and is prevented

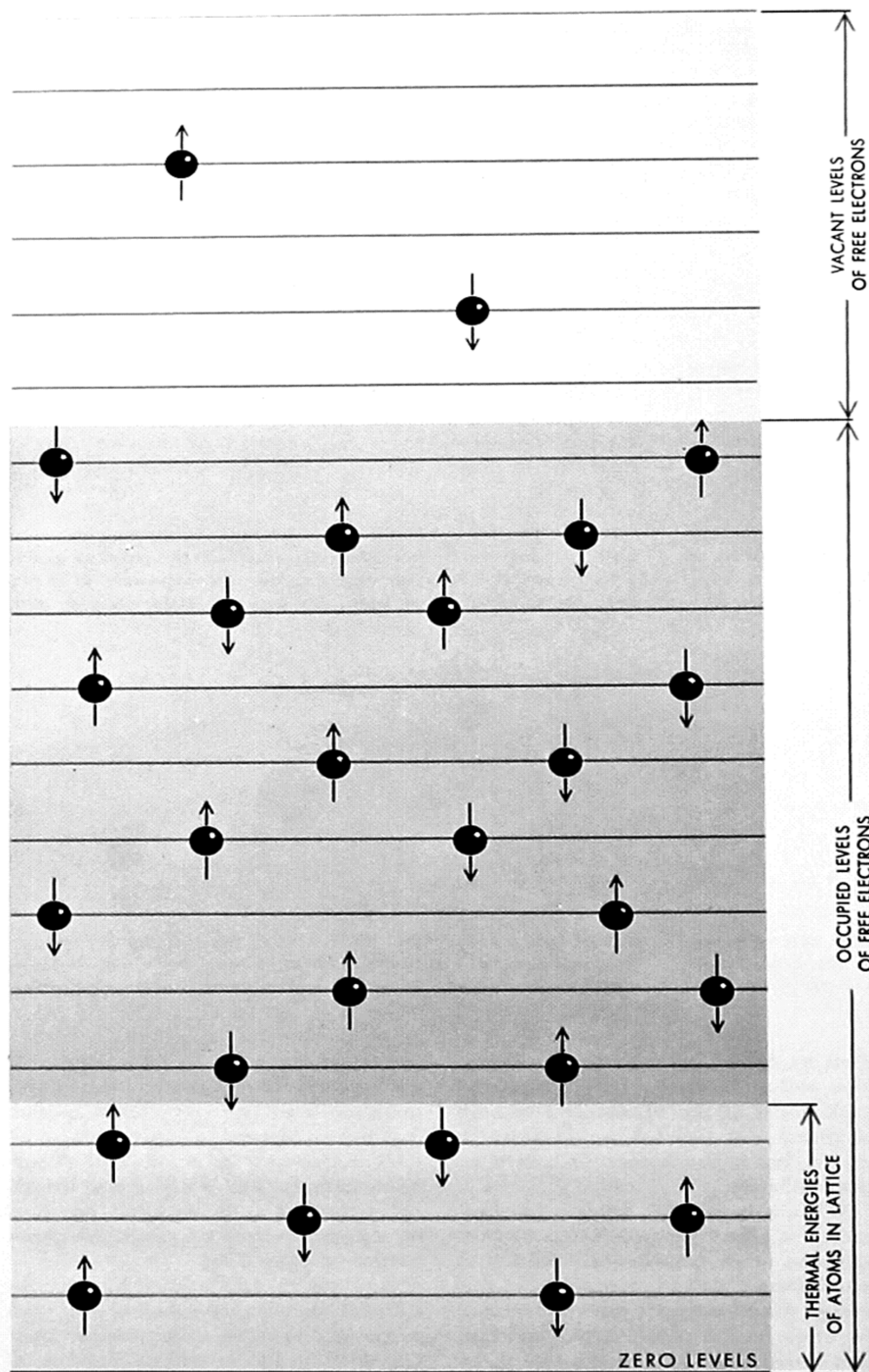
from escaping from it by surface forces. This picture leads to a very satisfactory explanation of the phenomena of both electrical and thermal conductivity in metals, but it also leads to one very serious difficulty.

When a material is heated, the heat energy it absorbs increases the thermal agitation of its constituent particles. It was expected that in metals part of this energy would intensify the vibrations of the atoms in the crystal lattice and that part would increase the velocities of the free electrons. However, studies on heated metals showed that this is not so; all the heat absorbed by a metal goes to increase its lattice vibrations.

How can this possibly be? The answer is that the motion of the free electrons is quantized, even though they are no longer restricted to atomic orbits. Atomic electrons are generally restricted to one of a few hundred quantum orbits, but

free electrons have literally billions of quantum levels available to them. These quantum levels are very closely spaced, forming, so to speak, a ladder with an almost infinitely large number of closely spaced rungs. The lower-energy rungs, being the first to be occupied, are filled to capacity by free electrons [see illustration on page 526].

Any gas in which particles are distributed among quantum levels in such a way is known as a "degenerate gas." Considering the enormous number of free electrons in metals, we realize that the energy spectrum of this gas must range from rather small values, for the lowest quantum-levels, to extremely high values, for the uppermost occupied levels. The energy of the fastest of the free electrons greatly exceeds the mean energy of lattice vibrations at normal temperatures. In fact, the energy of the free electrons is so much higher than the



FREE ELECTRONS in a metal are forced by the exclusion principle to occupy a series of energy levels arranged like rungs on a ladder. Only two electrons of opposite spin (*arrows indicate spin orientation*) are permitted on each level. The energy levels of free electrons (*gray areas*) are much higher than the thermal vibration energies (*colored area*) of the atoms forming the metal's crystal lattice. When the metal is heated, thermal lattice vibration increases. Because of their higher energies, the motion of free electrons does not increase.

mean energy of thermal lattice-motion that an elevated temperature will shake the atomic lattice apart (that is, melt the metal) long before it begins to affect the motion of free electrons.

We encounter another degenerate gas in the interiors of small, faint stars—the so-called white dwarfs, which represent the late stages of stellar evolution. The matter within these stars is completely dissociated and compressed to unimaginable densities. All the atoms are stripped down to bare nuclei; all their electrons are torn from their customary orbits, forming a free-electron gas. This gas is quantized, and the electrons are distributed according to the Pauli principle among a gigantic number of quantum states. Since the electrons possess much higher energies than the average thermal energy inside these stars, the gas exerts a tremendous pressure which prevents them from collapsing [see “Dying Stars,” by Jesse L. Greenstein; *SCIENTIFIC AMERICAN* Offprint 216].

Exclusion and Antimatter

We find another extremely important application of the Pauli principle in the theory, propounded 29 years ago by P. A. M. Dirac of the University of Cambridge, that predicted the discovery of the positive electron, or positron. Dirac's relativistic quantum-mechanics postulates that each fundamental particle can exist in one of two different physical states: the “ordinary” state, in which we find both atomic and free electrons, and the “extraordinary” state, in which the particles have negative (less than zero) energy. From the theory of relativity we know that particles with negative energy also possess what is known as negative mass, *i.e.*, the property of being accelerated in a direction opposite to that of an acting force. Particles in the extraordinary state had never been observed, but the perfectly consistent arguments of Dirac's theory led to the inevitable conclusion that not only does the extraordinary or negative-energy state exist, but also that because of its lower energy it is more stable than the ordinary state. It follows that all particles in the universe tend to change their state from the ordinary to the extraordinary one—which would turn the physical properties of matter into an unbelievable mess!

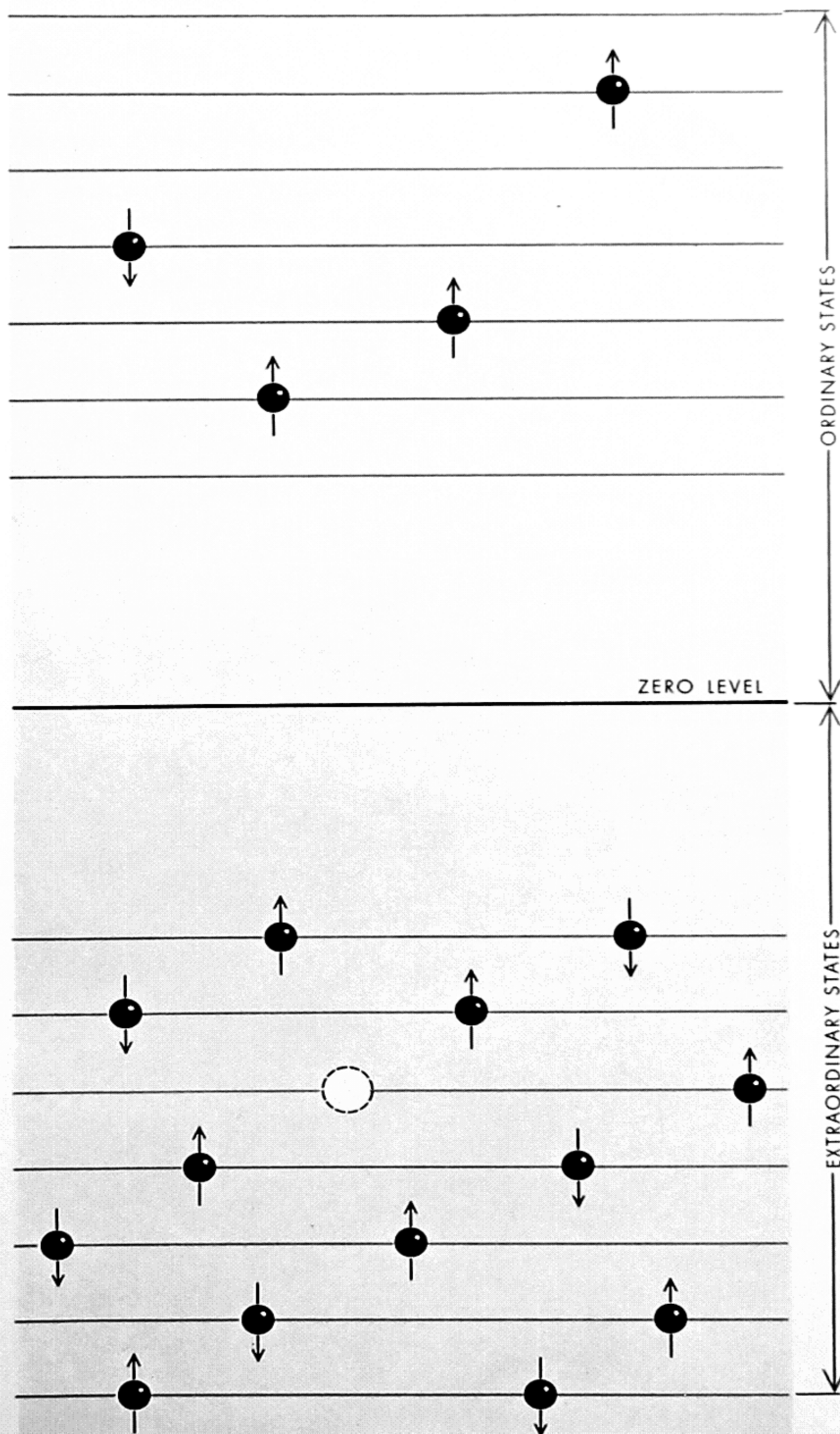
Why doesn't this happen? Dirac's answer was just as fantastic as his paradoxical particles. He postulated that quantum levels representing the extraordinary states of elementary particles

are already completely filled, and that the ordinary particles with which we deal in our everyday physical world are simply the excess that cannot be accommodated at the lower energy-levels because of the Pauli principle! According to this viewpoint, a vacuum is not really an empty space; on the contrary, it is a sea of tightly packed particles of negative mass. We are not aware of these extraordinary particles because they are distributed uniformly through space, putting us in somewhat the same position as a deep-sea fish which, though it is surrounded by water on all sides, may not be aware that it is floating in a physical medium. The ordinary particles that we can observe physically are those that are prevented by the Pauli principle from giving up their energy and falling into the negative energy-levels of the extraordinary state; ordinary particles must maintain their positive mass and all their familiar physical properties.

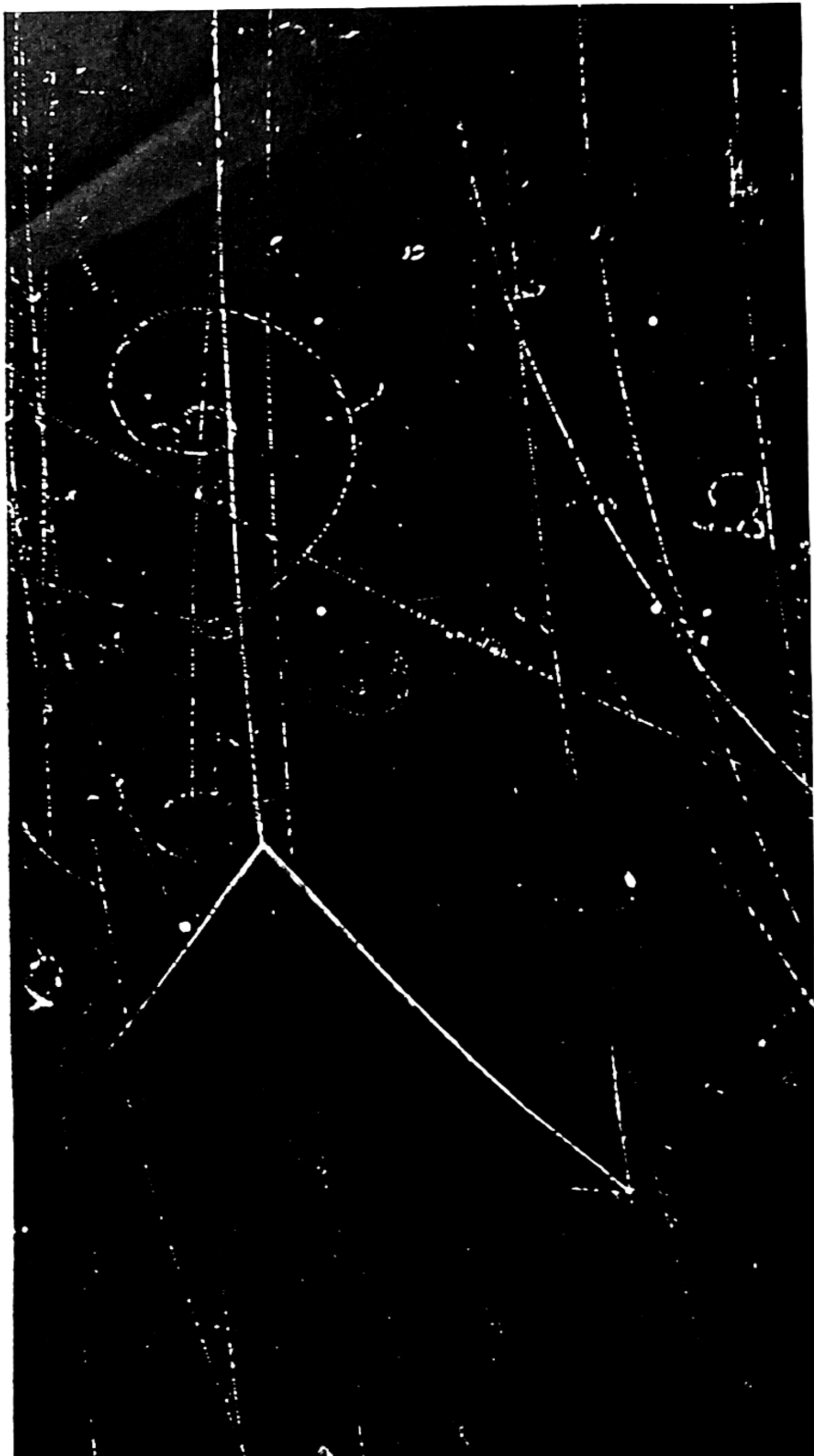
Unbelievable though Dirac's theory seemed (and it was strongly criticized by all theoretical physicists when it was first published), it led to one very important conclusion. If one of these extraordinary particles of negative mass is absent for some reason, it leaves a "hole" in the observable continuous distribution of particles. It is easy to see that we should be able to observe this hole, just as our fish is able to observe an air bubble (the absence of water in a small region) rising toward the surface of the ocean. What we perceive in the absence of negative mass is the presence of an equal amount of positive mass, just as we perceive the absence of charge in an electric field as the presence of an equal charge of opposite sign. In other words, a hole in the continuous distribution of extraordinary particles will seem to be an "antiparticle"—a particle of ordinary mass and opposite charge [see "Antimatter," by Geoffrey Burbidge and Fred Hoyle; *SCIENTIFIC AMERICAN* Offprint 202]. In fact, the discovery of positive electrons, the properties of which coincided exactly with the predicted properties of these holes, fully confirmed Dirac's conclusions a few years after he had published his paper.

The Quantized Nucleus

More recent developments in nuclear physics have extended the domain of the Pauli principle still further, this time into the nucleus itself. It has been found that protons and neutrons, the two main components of the nucleus, behave in many respects like the electrons that form the



UNOBSERVABLE ELECTRONS, which have negative energy and negative mass, are distributed evenly throughout space. Observable "ordinary" electrons are those that the exclusion principle prevents from falling into the "extraordinary" negative-energy states. "Holes" (open circle) in the continuous distribution of extraordinary electrons can be detected as positrons, particles of antimatter having positive mass and positive charge.

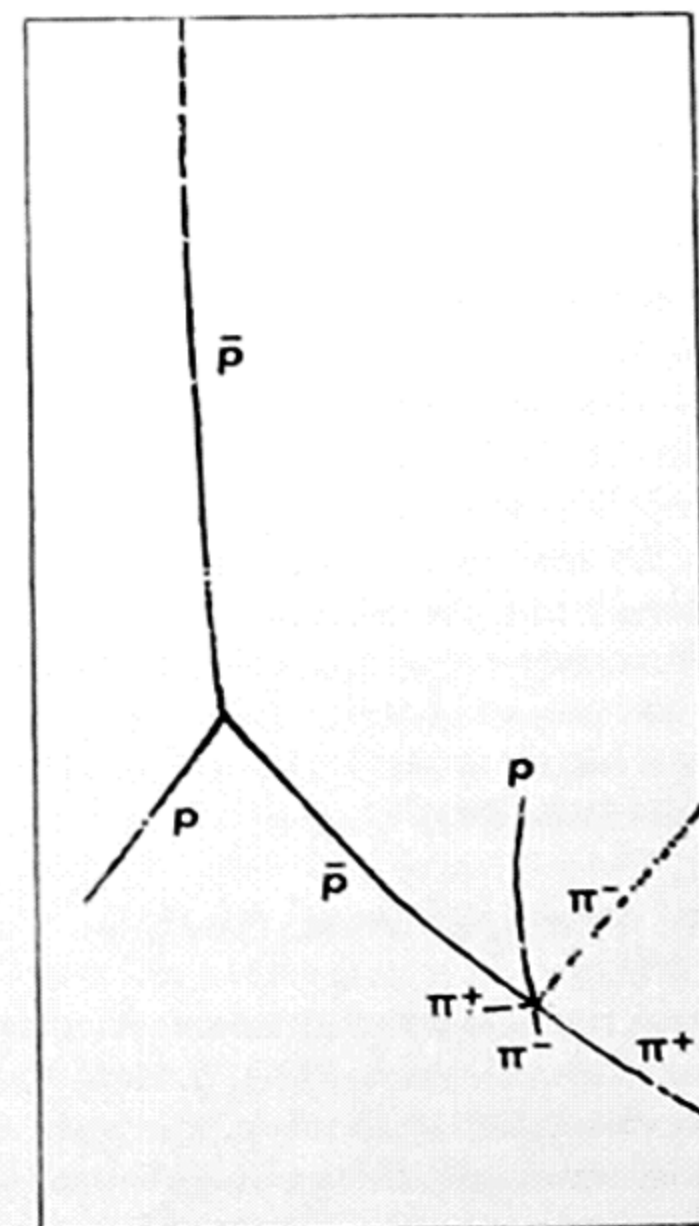


PARTICLE OF ANTIMATTER leaves its tracks in the bubble-chamber photograph at left, made by the research groups of Wilson Powell and Emilio Segrè at the University of California. The drawing at right traces the path of the particle in the photograph. A

outer shells of the atom. Like electrons, protons and neutrons possess magnetic moment and may be considered little spheres spinning in one direction or another; similarly the motion of particles within the nucleus is quantized so that only two particles of the same kind (with opposite spins) are permitted on each quantum orbit. But because the forces within the nucleus act rather differently from the electrical forces that hold the atom's electron shells in place, the pattern of nuclear energy-levels is also rather different.

It was first shown by observation and later explained by theory that consecutive shells inside the nucleus can accommodate 2, 8, 14, 20, 28, 50, 82 and 126 particles of either kind. These are the so-called magic numbers of the nucleus; they represent a complete analogy to the numbers of electrons that can be contained in the consecutive shells outside the nucleus. Just as atomic structure would be quite different if the Pauli principle did not apply to electrons, so nuclear structure would be quite different if the principle did not apply to protons and neutrons.

The discovery of positive electrons suggested that protons and neutrons might also occupy extraordinary states of



high-energy antiproton (\bar{p}) enters the chamber at upper left, and collides with a proton (p), which recoils toward the left side of the picture. The antiproton ricochets to the right and is annihilated in a carbon nucleus, producing four pi mesons (π) and a proton.

negative energy. Physicists accordingly began to look for negatively charged protons (antiprotons) corresponding to holes in the continuous distribution of extraordinary positive protons of negative mass. And they instigated a search for the more elusive antineutrons, which represent holes in the continuum of extraordinary neutrons. They have found them both; particle-accelerator experiments in recent years have brilliantly confirmed Dirac's predictions of extraordinary energy-states governed by the Pauli principle.

Wave Mechanics

After a review of the application of the Pauli principle to atomic and nuclear structure, to metallic conduction and to stellar interiors, and finally to the theory of antiparticles, we might feel that we have at last grasped the principle's physical meaning. But we are mistaken. Full comprehension requires that we understand the principle in terms of modern quantum-theory. Up to this point our review has employed the old-fashioned metaphor that pictures electrons and protons as tiny, electrically charged spheres spinning on their axes and moving along circular or elliptical orbits. But we now know that this picture, which is based on our everyday experience with large bodies such as flying bullets and spinning tops, is naive and basically incorrect when it is applied to the atomic world. When we speak about atomic particles, we must abandon our mechanical cause-and-effect description of phenomena and use the cumbersome though exact language of wave mechanics.

In wave mechanics the different orbits in Bohr's hydrogen atom correspond to different probabilities of finding the electron at various locations within the atom. The "vibration modes" of these probability functions [see illustration at right] tell us all that can be told about the motion of atomic particles. In wave mechanics the old Bohr statement that "the electron moves along the first, second or third quantum orbit" is paraphrased to say that "the first, second or third vibration mode is excited within the atom." The jump of an electron from one quantum orbit to another is interpreted as the dying-out of one vibration mode and the simultaneous appearance of another. Thus the entire theory of electron motion within atoms takes on a physical meaning closer to the theory of organ pipes, violin strings or drumheads than to the theory of planetary motion around the sun. In these terms the Pauli principle becomes equivalent to the

statement that each particular vibration mode can either be excited or not, just as we can strike two different keys on a piano simultaneously but cannot strike the same key twice at the same time!

Pauli first stated his exclusion principle in terms of the mathematical "symmetry" of probability functions. A probability function is mathematically symmetrical if its variables (which, in the case of the electron, take account of both orbital motion and spin direction) can be interchanged without changing its sign (plus or minus). It can be easily shown that the exclusion principle is equivalent to the following statement: The probability function describing the motion of electrons within an atom is antisymmetrical with respect to any electron pair.

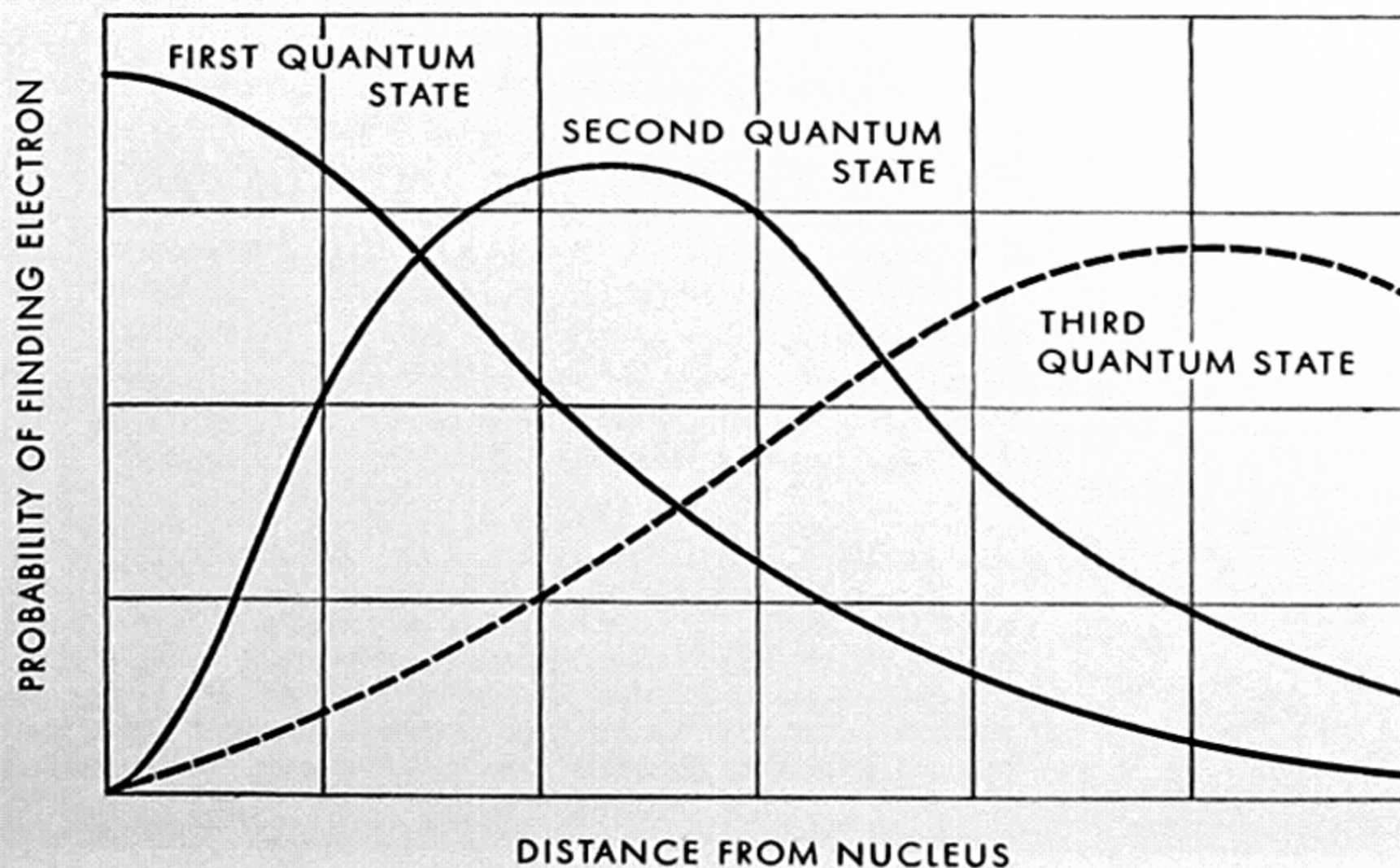
If we mathematically transpose the positions of two electrons moving along the same atomic orbit and spinning in the same direction, their probability function does not change in value. The Pauli principle, however, tells us that because their probability function is antisymmetrical, it must change sign as the result of such a transposition. But what function can change sign without changing in value? The only quantity that satisfies these conditions is zero. Plus zero equals minus zero; changing the sign of zero does not change the value of the function in any way. Thus it follows that the situation where two electrons with the same spin move along the same orbit has a probability function of zero.

In other words, the situation is impossible. And this is exactly the way in which Pauli originally stated his exclusion principle in 1925.

Non-Pauli Particles

Like Dirac's antimatter theory, Pauli's exclusion principle was first postulated for electrons, then successfully applied to protons and neutrons. Since then physicists have raised the question of whether or not the exclusion principle applies only to elementary particles such as protons and electrons or whether it might also apply to whole nuclei or even to whole atoms. We can answer this question by theoretically applying the Pauli principle to two helium nuclei (alpha particles), which are composite particles composed of two protons and two neutrons. What happens to the probability function describing the motion of two alpha particles if we exchange their positions?

We can carry out such an exchange in four mathematical steps. First we transpose one proton belonging to the first alpha particle with one of the protons belonging to the second. Then we transpose the other pair of protons and finally each of the two pairs of neutrons. Since protons and neutrons obey the Pauli principle, each transposition changes the sign of the probability function from plus to minus or from minus to plus, but at the end of four steps the function has the same sign as before. Thus we conclude



WAVE FUNCTIONS express the probability of finding an electron in a particular quantum state and at a given distance from the nucleus. These functions describe the energy and motion of the electron far more accurately than do the orbit concepts of classical theory.

that the probability function describing two alpha particles is symmetrical, or in other words that alpha particles considered as individual units are not subject to the Pauli principle.

But not all composite particles are exempt from the Pauli principle. The transposition of two tritons (nuclei of hydrogen 3, or tritium, containing one proton and two neutrons) does result in a change of sign, so we must conclude that tritons are subject to the Pauli principle. Furthermore, we find that atoms such as those of oxygen (eight protons, eight neutrons and eight electrons), containing an even number of elementary particles, are not subject to the Pauli principle; and that atoms such as those of nitrogen (seven protons, seven neutrons and seven electrons), containing an odd number of particles, must obey it. This curious even-odd discrimination

determines the statistical behavior of particles: The even-numbered or "non-Pauli" particles obey the laws of Bose-Einstein statistics and hence are called bosons; the odd-numbered or "Pauli" particles obey Fermi-Dirac statistics and are called fermions.

Until about two decades ago electrons, protons and neutrons were the only known elementary particles in nature. Later studies have revealed the existence of a large number of particles with a claim to elementarity. We now have three kinds of hyperons (unstable particles heavier than protons), three kinds of mesons (unstable particles with a mass somewhere between that of the proton and the electron) and neutrinos (with practically no mass at all). Although experimental studies of these new particles reveal new and exciting facts about them almost every month,

theoretical progress in understanding their properties is almost at a standstill. We do not know why they have the masses they do; we do not know why they transform into another the way they do; we do not know anything! The one concept that stands like the Rock of Gibraltar in our sea of confusion is the Pauli principle, which separates the fermions (such as neutrinos and mu mesons) from the bosons (such as pi mesons and lambda particles).

Is the fact that some of these particles are "Pauli" and some "non-Pauli" connected with their internal structure, as it is in the case of alpha particles and tritons? Are some of these particles elementary, while others are not? We do not know. Let us hope that sometime soon the fog enveloping these questions will be dispersed by somebody's ingenious idea.

The Author

GEORGE GAMOW was born in Odessa, Russia, in 1904. As a graduate student of physics at the University of Leningrad in 1928, he found himself in hot water because "the research topic proposed by my professor was so boring that the work hardly made any progress." That summer Gamow took some courses at the University of Göttingen. While there he devised a quantum theory of radioactivity which shed new light on the structure of the atomic nucleus. Instead of returning to Leningrad he accepted an invitation from Niels Bohr, who had learned of his work, to spend a year at the University of Copenhagen on a Fellowship provided by the Carlsberg brewing firm. The following year he worked with Ernest Rutherford at the University of Cambridge on a Rockefeller fellowship. At Cambridge he wrote his first book, entitled *Constitution of Atomic Nuclei and Radioactivity*. After two more years of teaching at Lenin-

grad, Gamow attended the International Solvay Congress on Physics in Brussels and decided not to return to the U.S.S.R. In 1934 he accepted a professorship at George Washington University; in 1956 he became professor of physics at the University of Colorado. After Gamow came to the U. S., his interests shifted from pure nuclear physics to its applications in cosmology, and later to fundamental problems of biology. A prolific popularizer of science, Gamow was awarded the Kalinga Prize in 1956 for his interpretation of science for the layman. He has published a dozen books in 23 languages, with three more going to press.

Bibliography

EXCLUSION PRINCIPLE, LORENTZ GROUP AND REFLECTION OF SPACE-TIME AND CHARGE. W. Pauli in *Development of Physics*, edited by Niels Bohr, pages 30-51; McGraw-Hill Book Company, 1955.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

MOLECULAR MOTIONS

by B. J. Alder and T. E. Wainwright

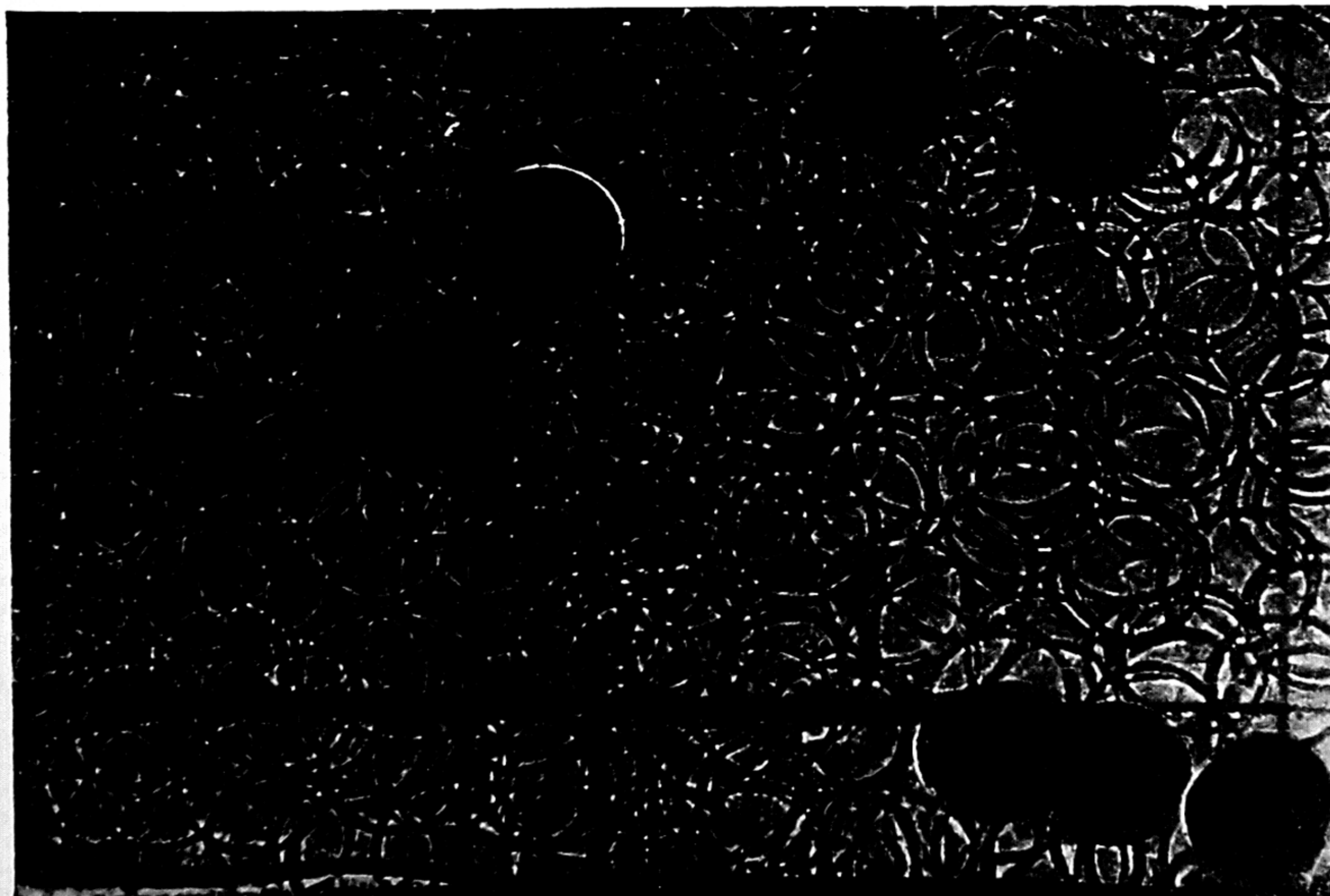
One of the aims of molecular physics is to account for the bulk properties of matter in terms of the behavior of its particles. High speed computers are helping physicists realize this goal.

During the 19th century, as evidence in favor of the atomic theory mounted, an ancient hope of science began to bear fruit. As long ago as the first century B.C. Lucretius had proposed not only that matter is composed of tiny particles called atoms, but also that the behavior of these par-

ticles is the key to understanding the properties of bulk matter. For centuries this idea remained simply an interesting hypothesis. Then Isaac Newton set forth laws of motion from which the behavior of atoms might be calculated. At the same time a number of investigators were making quantitative observations

on the gross properties of matter. The stage was set for an attempt to realize Lucretius' dream.

The earliest tries were mostly unsuccessful. But in 1739 Daniel Bernoulli succeeded in proving that the product of the pressure and the volume of a gas is proportional to the average kinetic



PHYSICAL MODEL of a molecular system consists of gelatin balls suspended in a tank of liquid. The tank is shaken to generate

typical patterns of separations between balls. In the experiment depicted here separations were measured for the seven black balls.

energy of its atoms, provided that the atoms do not interact. His result is in fact valid for gases at very low density, where interactions between atoms or molecules are rare. It was not until the 19th century, however, that the program got under way on a grand scale, with the work of such men as Rudolf Clausius, James Clerk Maxwell and Ludwig Boltzmann. The task they faced was immense. Their predecessors had dealt with such problems as computing the orderly motions of a few planets under the gravitational attraction of the sun. These men were concerned with millions of particles, colliding with one another and darting about in all directions. To follow the trajectory of any individual atom in a piece of matter it would be necessary to know the position at all times of every other particle close enough to exert a force on it. The total force on the atom could then be computed from instant to instant, and, assuming that Newton's laws applied, its motion could be calculated.

In practice such a detailed calculation was hopelessly complicated for a dozen particles, let alone millions. Fortunately it was not necessary. The bulk properties of matter depend upon the average behavior of many atoms, and not upon the detailed motion of each. Therefore statistical methods could be used and a new branch of physics known as statistical mechanics grew up. The statistical approach fulfilled many of its founders' hopes. In some problems, however,

even this type of mechanics bogged down in cumbersome mathematics.

Today we are in a position to overcome some of the practical difficulties. Using high-speed computers we can perform calculations that were hitherto impossible. With the help of these machines we are moving a little closer to the ideal of understanding the properties of matter in terms of the mechanical behavior of its constituent particles.

In many applications of statistical mechanics it is possible to proceed with no knowledge of the velocities of individual atoms. We ask only how the atoms are distributed on the average in space. From this distribution alone many of the properties of the material can be calculated.

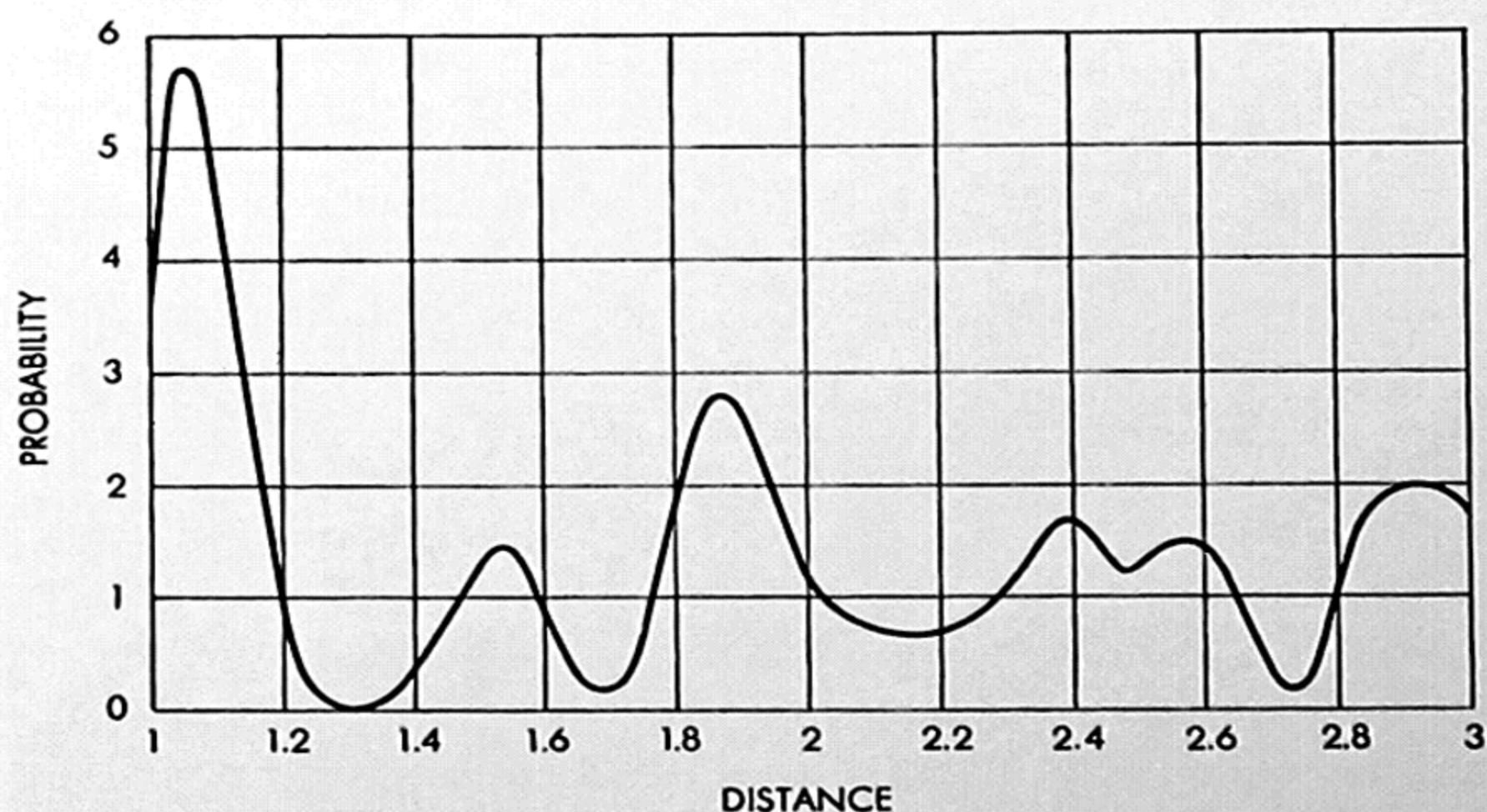
To appreciate what is meant by the spatial distributions, imagine that we have a microscope powerful enough to see the individual atoms or molecules in a sample of matter, and a camera fast enough to stop them in their rapid flight. A stereoscopic picture made with this arrangement shows how the particles are distributed at a given instant. We examine the particles in the snapshot in turn, measuring the distances between each one and all the others. The information is summarized on a graph, with distance of separation plotted along the horizontal axis, and the number of pairs at each distance along the vertical [see illustration below].

Suppose we make a number of snap-

shots in rapid succession, say 1,000 in a second. We plot the distances of separation in each picture and then obtain the average of all the curves. This graph represents the average distribution of pair separations for that second. If the system is in equilibrium, then the average distribution remains constant for all time.

Between each pair of molecules in a piece of matter there is a force that depends on the distance between them. Assuming that we know how the force varies with distance, we can, for example, use our average plot to compute the forces exerted on a typical molecule by its neighbors. This gives us the pressure of the system.

How do we actually find the distances of separation in a system of invisible particles darting about erratically at tremendous speeds? In some cases we can obtain the information experimentally. A beam of X-rays sent through a solid or a liquid is diffracted in a way that depends on the spacing between the particles of the substance. From the size and intensity of spots in the diffraction pattern we can deduce the distances of separation and the number of pairs at each separation. For a crystalline solid the pair-separation graph shows a number of distinct peaks and valleys, as in the illustration on the opposite page. The peaks indicate that the molecules tend to lie preferentially at certain specific distances from one another. This is what we should expect if the molecules are arranged in an orderly grid or lattice. As the temperature of the solid is lowered, the peaks in the plot become sharper, showing that the molecules vibrate less widely about their central positions in the lattice. With increasing temperature, on the other hand, the peaks grow broader. If the material is heated above its melting temperature, the peaks are still present at distances as small as a few molecular diameters. They disappear at large distances because the lattice structure has disintegrated, and there is no longer an ordering force between molecules at longer ranges.



DISTANCE-OF-SEPARATION PLOT for a crystalline solid is characterized by peaks and valleys. Peaks represent preferred distances of separation between the molecules of the crystal. Horizontal axis measures distance in molecular diameters. Vertical axis measures the relative probability that pairs of particles will lie at each distance. From this curve it is possible to calculate the actual numbers of pairs that are separated by each distance.

From the point of view of statistical mechanics we should like to be able to find the distribution of distance between molecules theoretically and to explain how physical conditions such as temperature and density influence the distribution. This would enable us, for example, to predict the pressure or energy of a substance from its temperature and density.

One way of approaching the problem

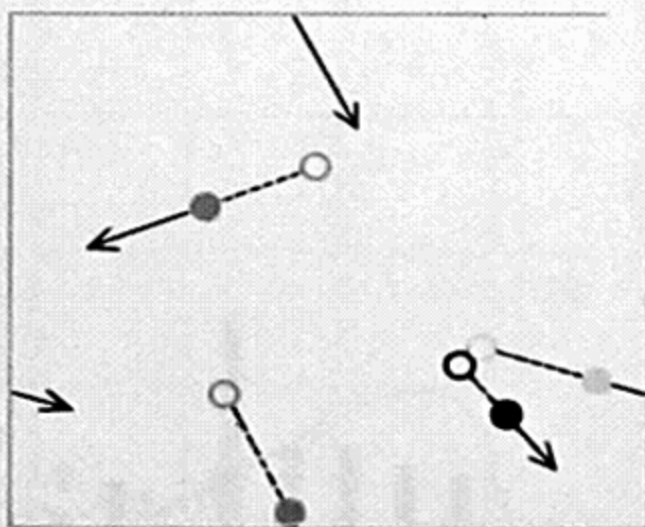
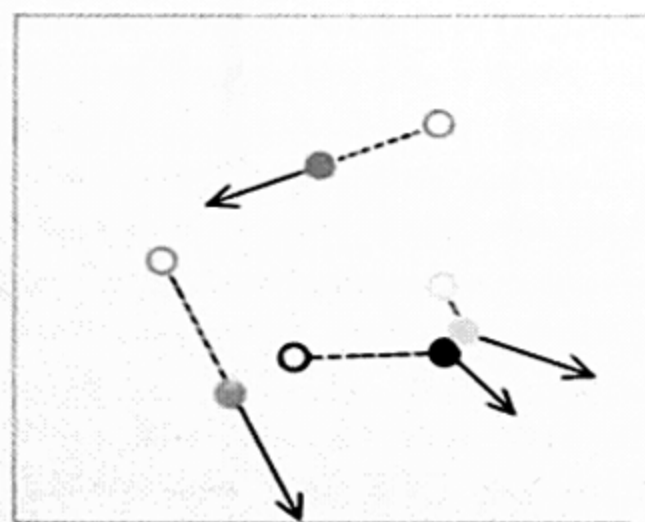
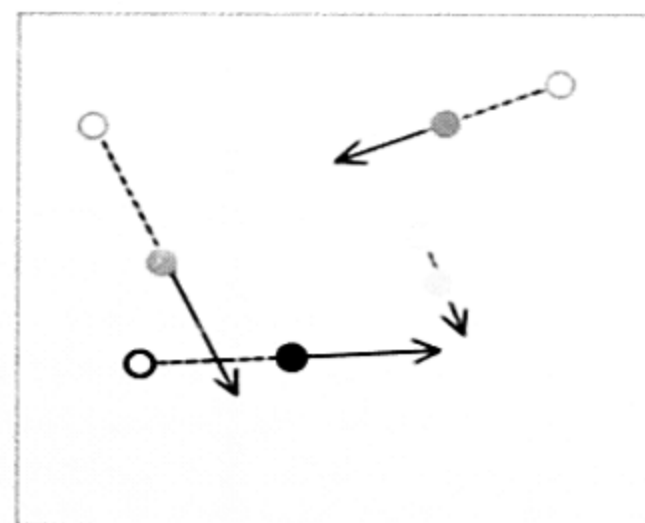
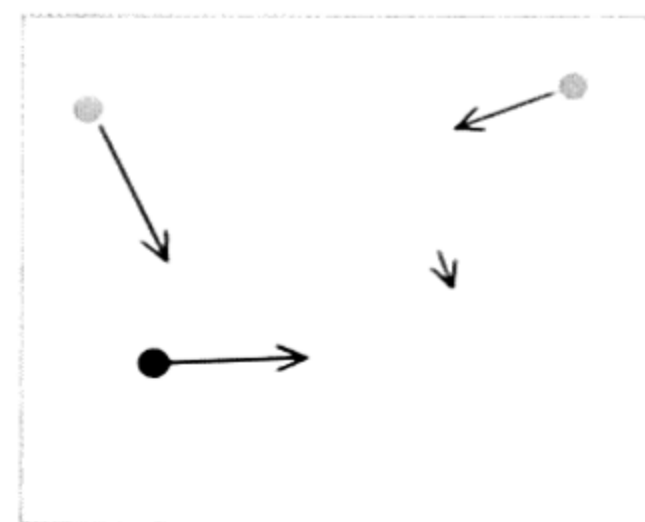
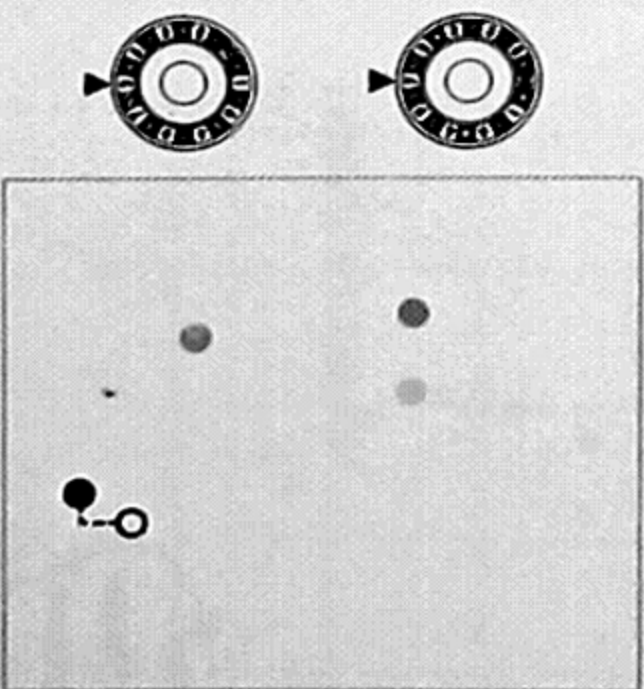
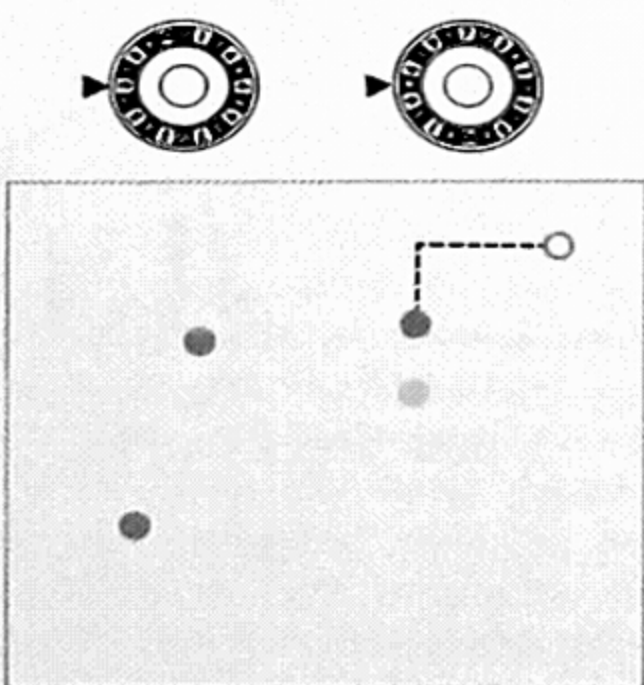
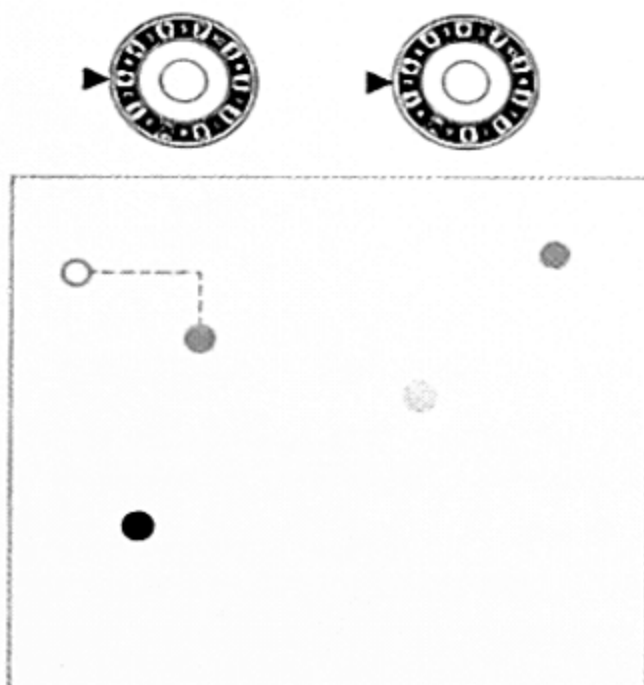
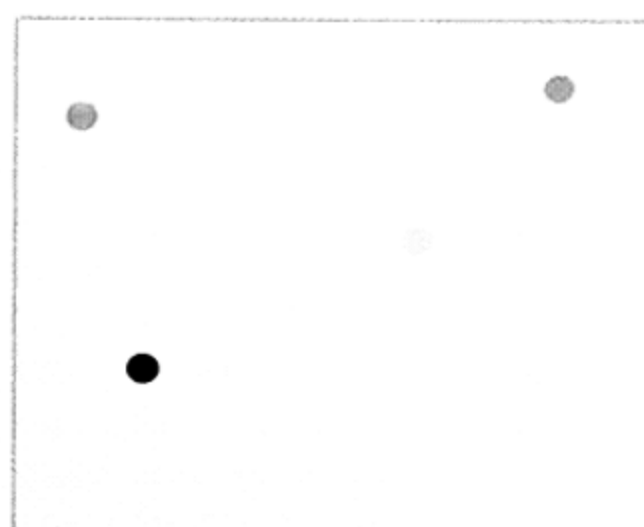
is to construct a mental model of a system of particles, usually assuming a simplified law of force between them, and to devise some means of "shaking" the system; that is, moving the particles about at random. From time to time we stop shaking and make a "snapshot," recording the distances of separation of all the pairs of particles. Then the snapshots are averaged.

It should be understood that the random moves of the particles in the model are not the same as the motions that carry molecules from one place to another in real matter. The moves are simply a device for generating possible spatial arrangements. Because the moves are artificial, the order in which the resulting configurations occur is also unrealistic. In a system in equilibrium, however, the order of snapshots does not matter. The average plot is the same regardless of the sequence in which the individual curves are considered.

A relatively easy model to deal with is one in which the particles are taken to be hard spheres, like marbles. There is no attractive force between them, and they repel each other only when they collide. Simple though it is, this model duplicates some of the properties of real systems surprisingly well, as we shall see later on.

To construct a hard-sphere system we choose imaginary marbles of a suitable size and place them in an imaginary box (usually a cubical one for the sake of convenience). There is not only no force of attraction between marbles, but also no force of gravity; each sphere stays just where it is put. The size of the box depends on the density assigned to our hypothetical material. If we have chosen a low enough density, the process is straightforward. The marbles can be put in at random, with little chance that any will overlap. Such an overlap is of course disallowed. Whenever it occurs, we must remove the last marble put in and try placing it elsewhere. The same thing is true after shaking: snapshots in which marbles overlap are ruled out. At sufficiently low densities, corresponding to the gaseous state in real materials, the frequency of overlaps is negligible. It is not too difficult to generate mathematically as many distance-of-separation patterns as we wish.

As the density is increased, however, that is, as the average distance between marbles gets smaller, it becomes harder to put in the marbles at random without running into interference. With each succeeding particle we must try



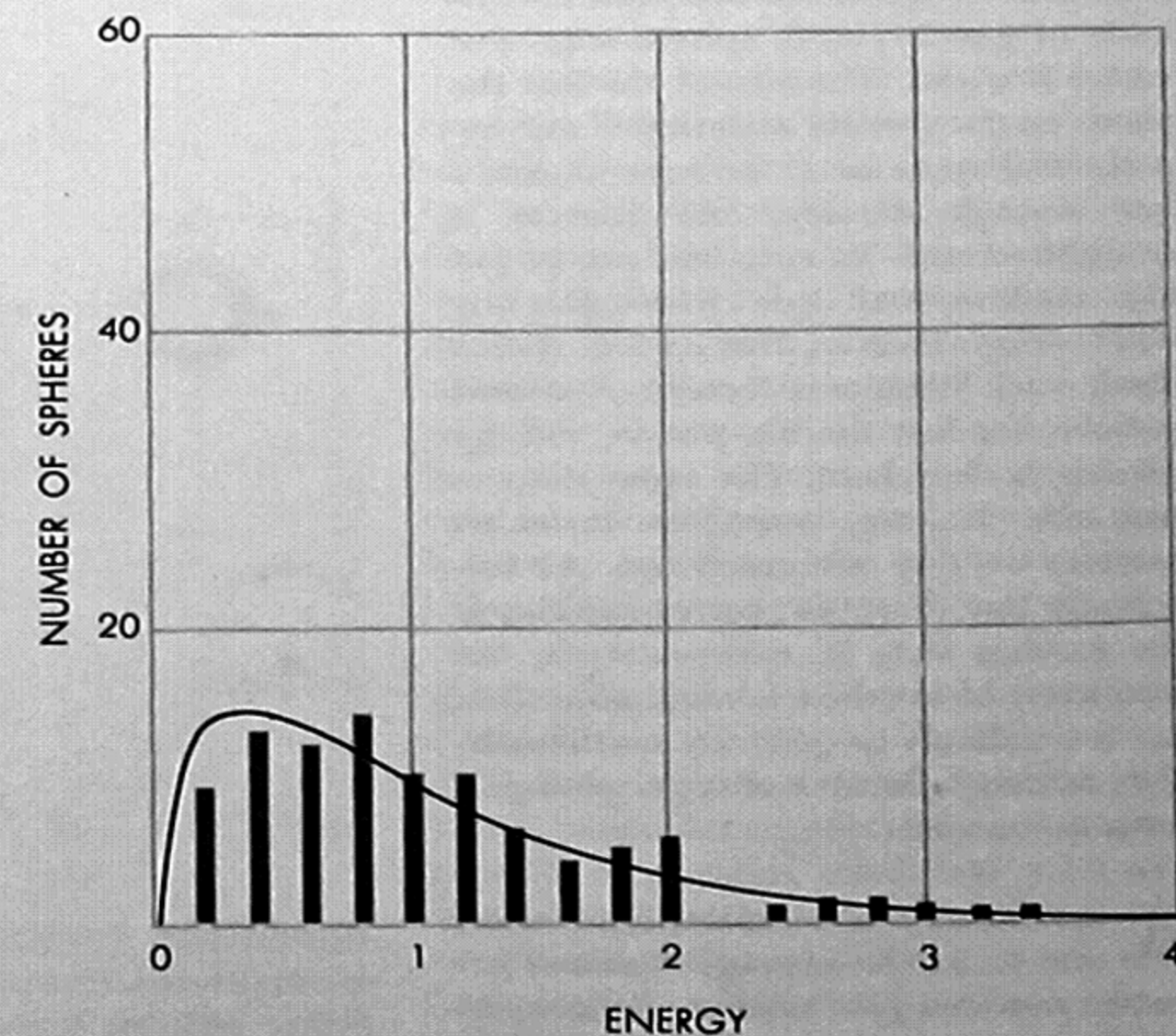
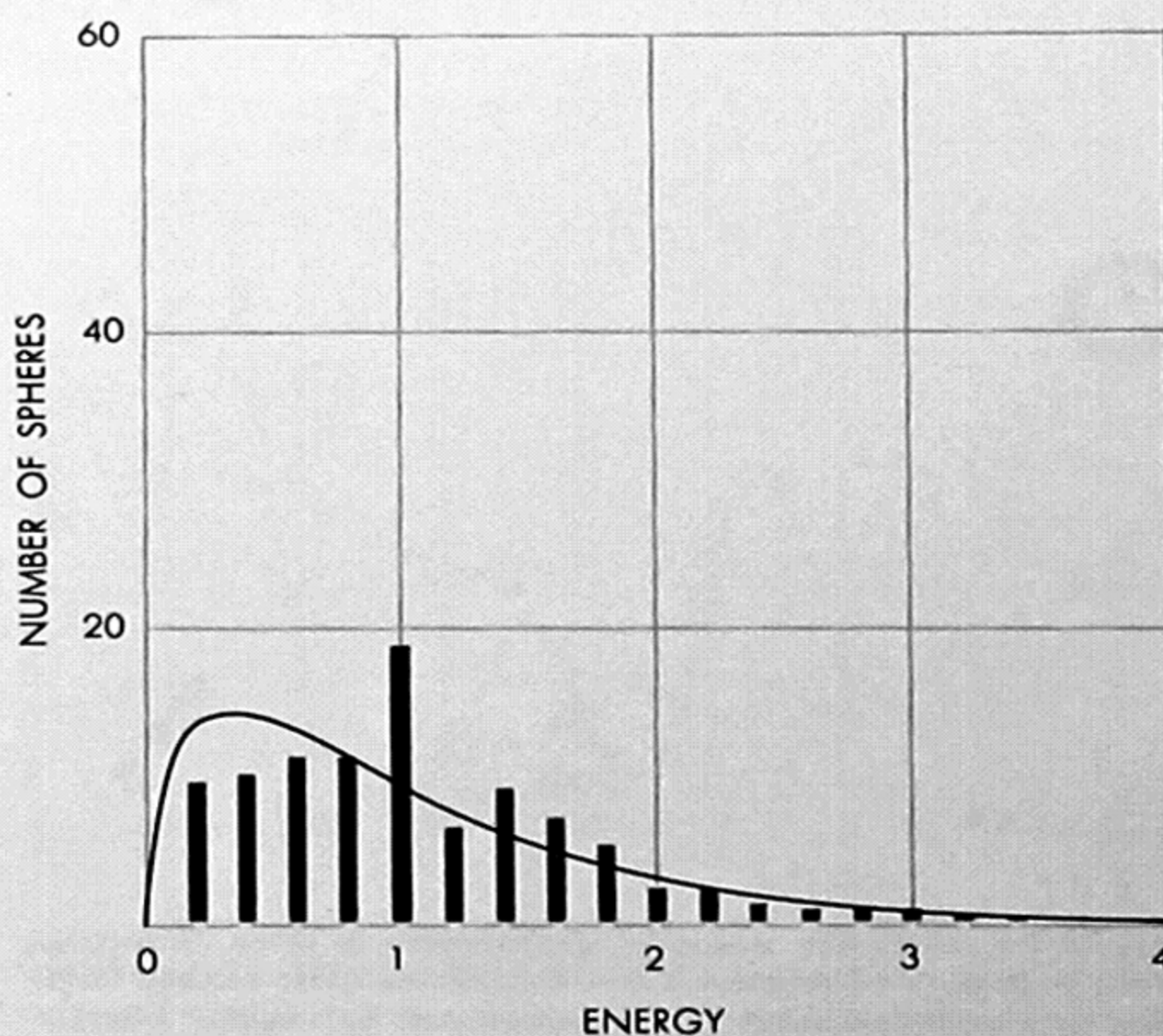
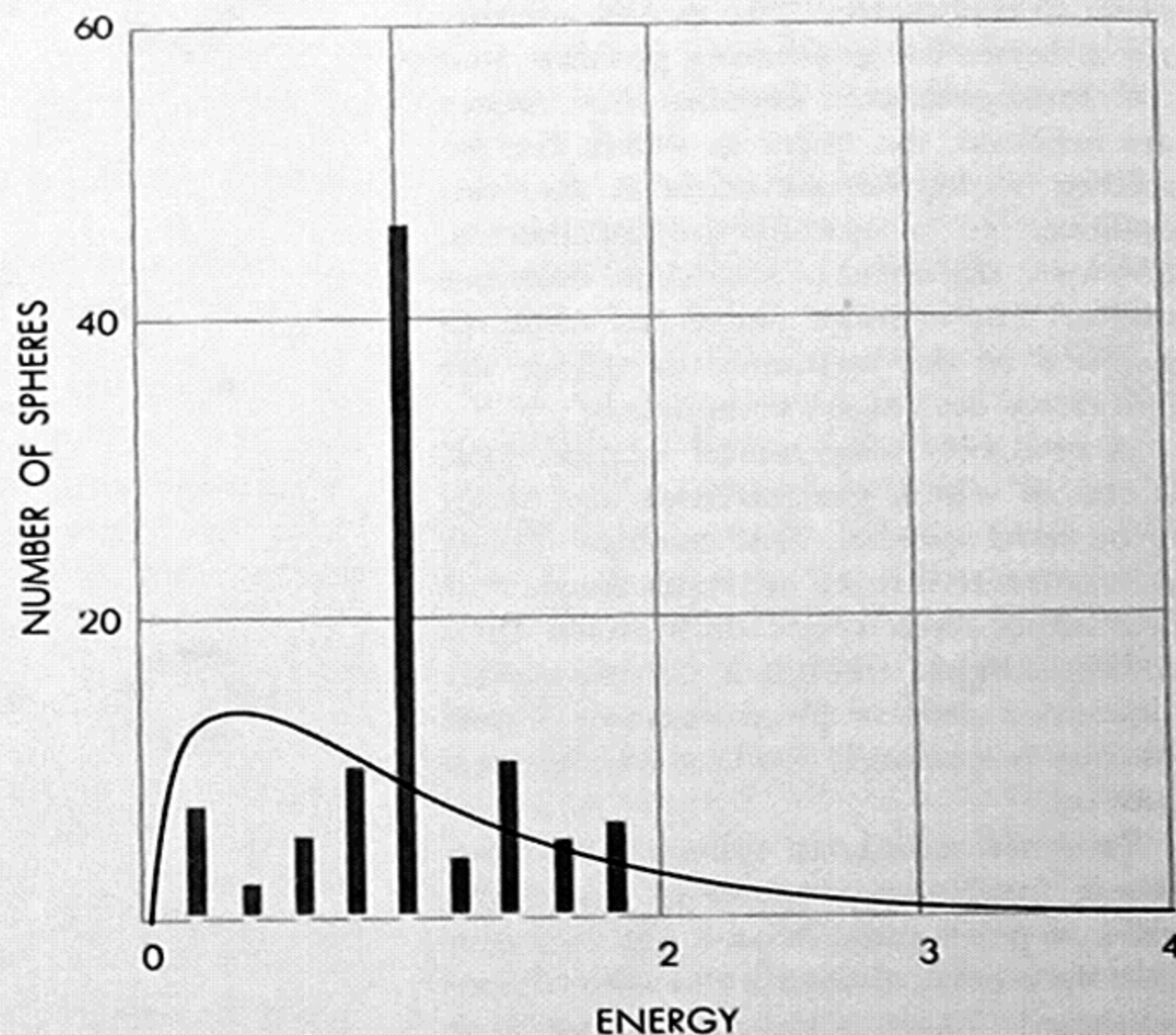
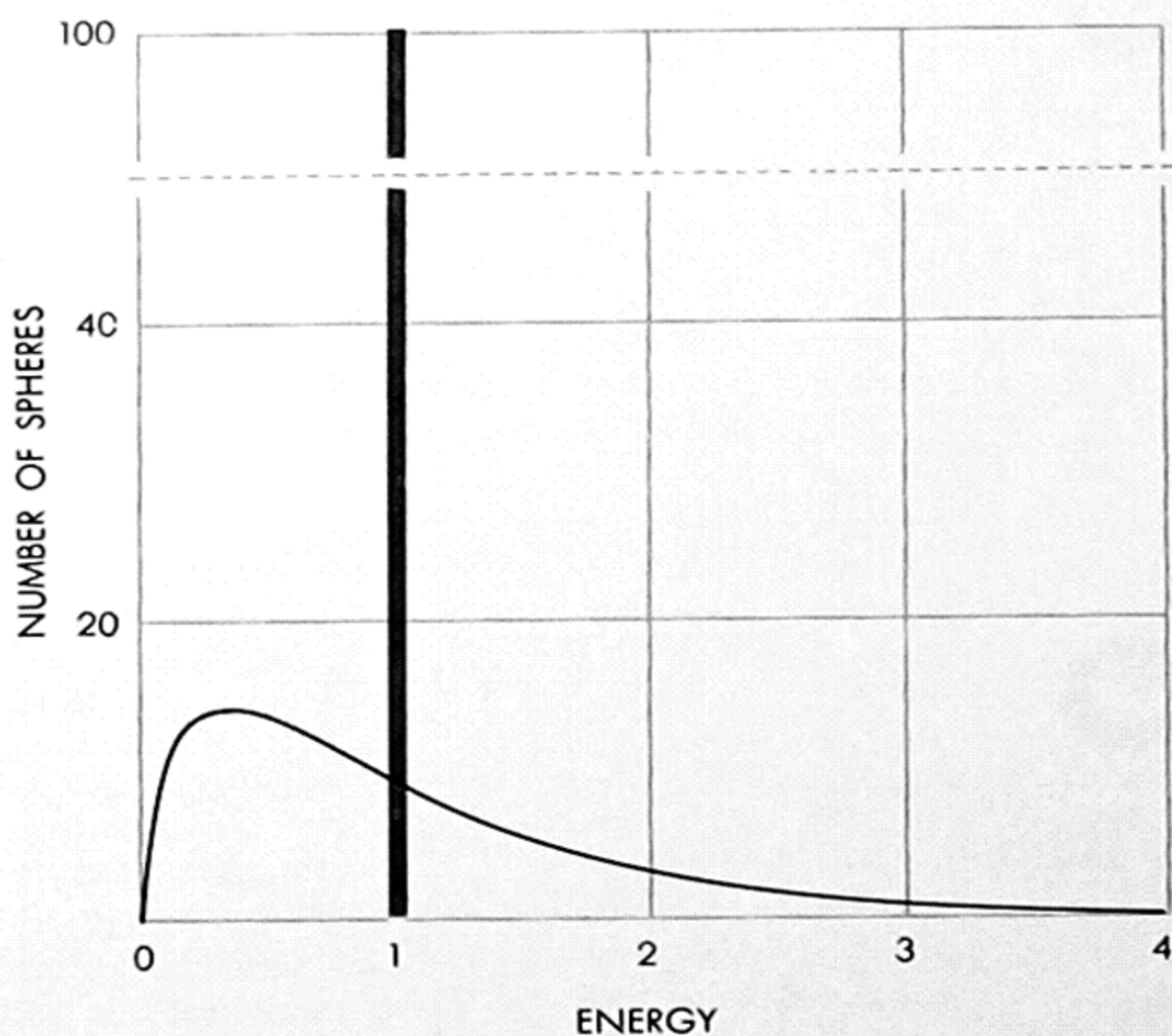
COMPUTER TECHNIQUES for calculating molecular distributions in three dimensions are illustrated schematically in these two-dimensional drawings. Monte Carlo method (*left*) moves particles according to a random-number table (represented by roulette wheels). Molecular-dynamical method (*right*) computes the actual trajectories. Open dots represent old positions of particles; solid dots, new positions. The particles are considered as being able to pass through the walls of the imaginary container and to re-enter on the opposite side.

more and more random positions before finding an empty spot. Eventually no more marbles can go in. The box is not completely full, but it is inefficiently packed. By rearranging the marbles already in it, we can make room for more. At this density matter begins to act like a liquid. In a liquid the spheres can rearrange themselves only if numbers of them move cooperatively. This situation makes the mathematics of shaking more cumbersome.

At still higher density, corresponding to that of a crystalline solid, the problem becomes simpler again. If the marbles are packed in an ordered array, and are touching one another, the box is completely full and shaking has no effect. The marbles cannot move at all. Now when the density is decreased slightly (by making the box bigger or the marbles smaller), the particles can rattle around their positions in the lattice, but cannot escape. This approximates the

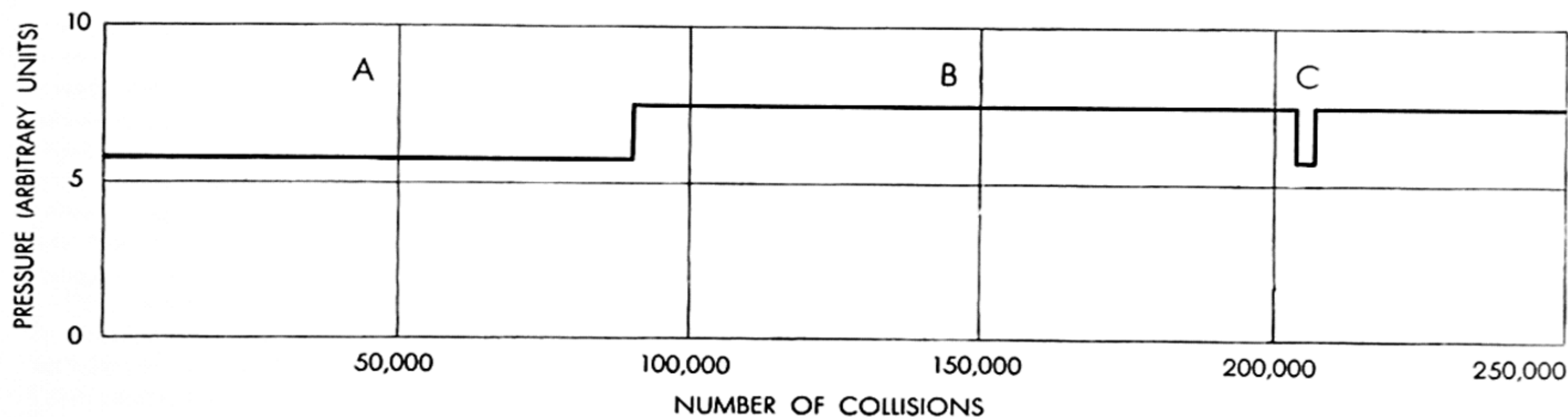
situation in a real crystal. If the marbles are shrunk still further, they will eventually be able to escape from the cages formed by their neighbors and trade places. The lattice disappears and the solid melts. But each particle can escape from its lattice position only if its neighbors move cooperatively in such a way as to leave it a wide enough path. Again the mathematics becomes harder to carry out.

One way around the computational



EQUILIBRIUM DISTRIBUTION of kinetic energies is rapidly attained by a system of 100 hard spheres, each having an energy of one unit (*top left*). Equilibrium pattern is represented by curve;

numbers of particles at various energies, by bars. Drawings at top right, bottom left and bottom right show distribution after 50, 100 and 200 collisions, or an average of one, two and four per particle.



PRESSURE-JUMPS in a hard-sphere system signal changes in phase between the fluid and solid state. Region A represents the low-pressure, solid phase, which lasted for about 100,000 collisions. Then

the pressure increased abruptly to that of a fluid (*region B*), continuing in this state for another 100,000 collisions. In region C it dropped briefly back to the solid value for another 3,000 collisions.

problem is to make a physical model and actually shake it. Some years ago Joel Hildebrand and his colleagues at the University of California performed just such an experiment. They suspended gelatin balls in a tank of liquid whose density was equal to that of the gelatin, so that the gravitational force was canceled by buoyancy. Placing the tank on a vibrating tray, they made a series of photographs from which the average distances of separation between the balls could be plotted. Their curve resembled plots made from X-ray studies of liquids.

No physical model, however, is altogether satisfactory. The chief problem is the difficulty of assuring that the shaking is really random.

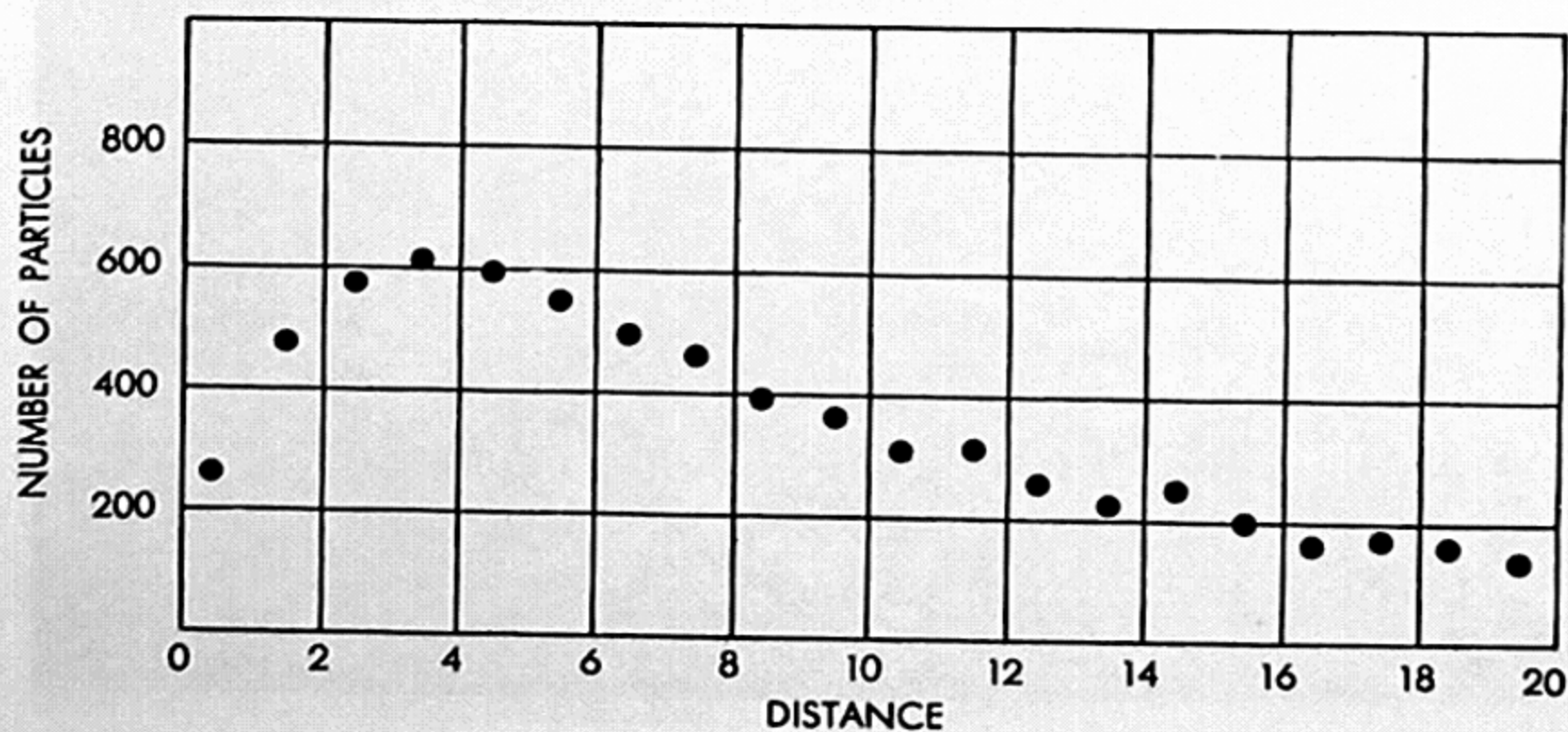
With the help of a fast computer the experiment can be "performed" much more neatly by purely mathematical means on an ideal system. This has been done by a group at the Los Alamos Scientific Laboratory. Each particle is represented by a set of three numbers, specifying the three-dimensional coordinates of its center (x , y and z). Having located all the particles by feeding the appropriate sets of numbers into the machine, we make our first "photograph," recording the distances of separation between all pairs. Now we proceed to "shake" the particles one by one, using the so-called Monte Carlo method. We choose a particle at random and displace it, that is, change its x , y and z coordinates. The amount of each change is decided by picking one of a series of random numbers generated by the machine. If the move is a legal one—that is, if the displaced particle does not overlap any of the others—we make a second snapshot, again recording all distances of separation. Then we displace a second particle, and so on.

Whenever a move turns out to be illegal, we return the displaced particle to its former position and use the corresponding distribution of distances a second time in the averaging. This procedure gives effect to the comparative probability of each distribution. The more often a random change in a distribution leads to an impossible configuration, the more probable is that distribution.

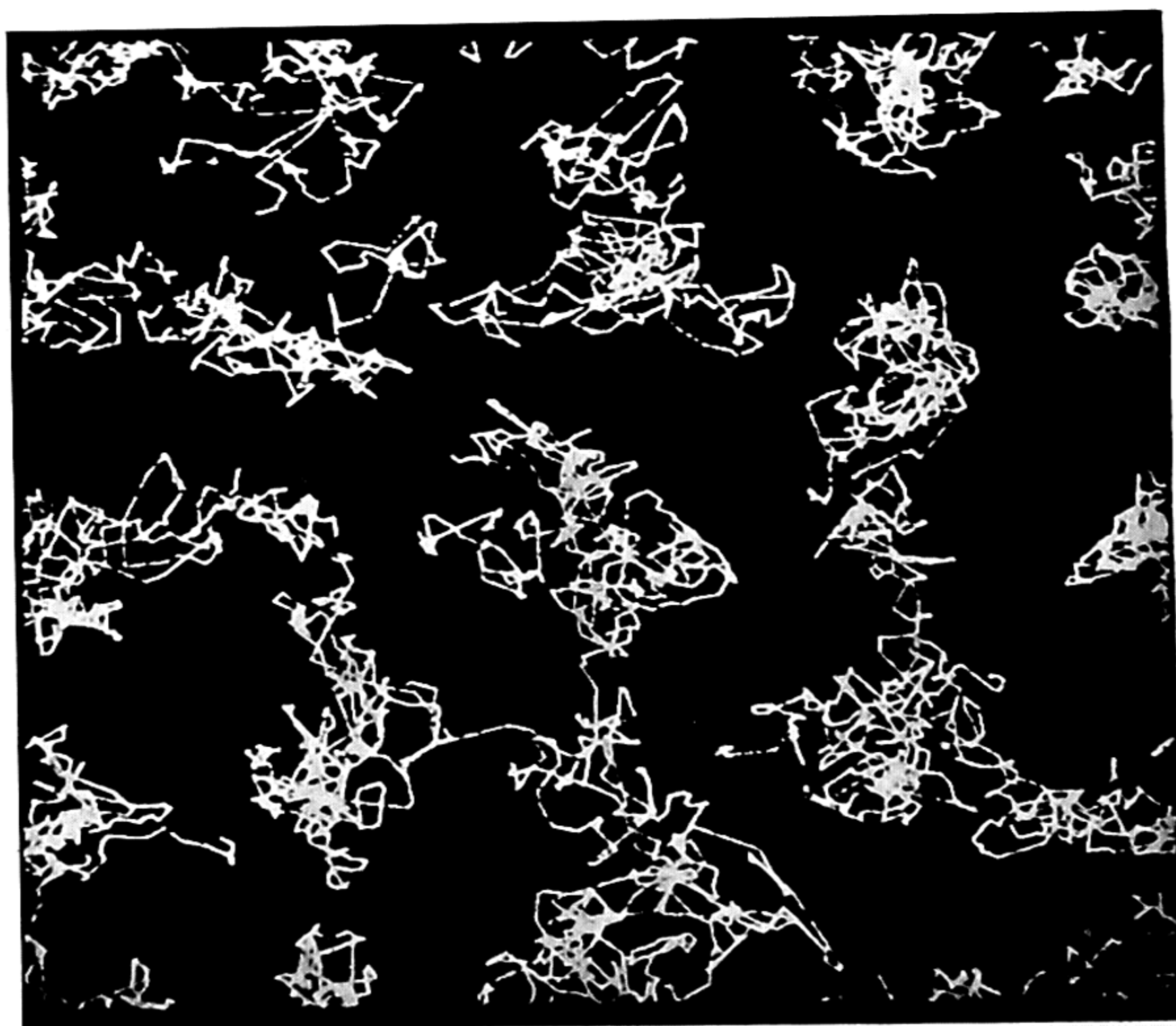
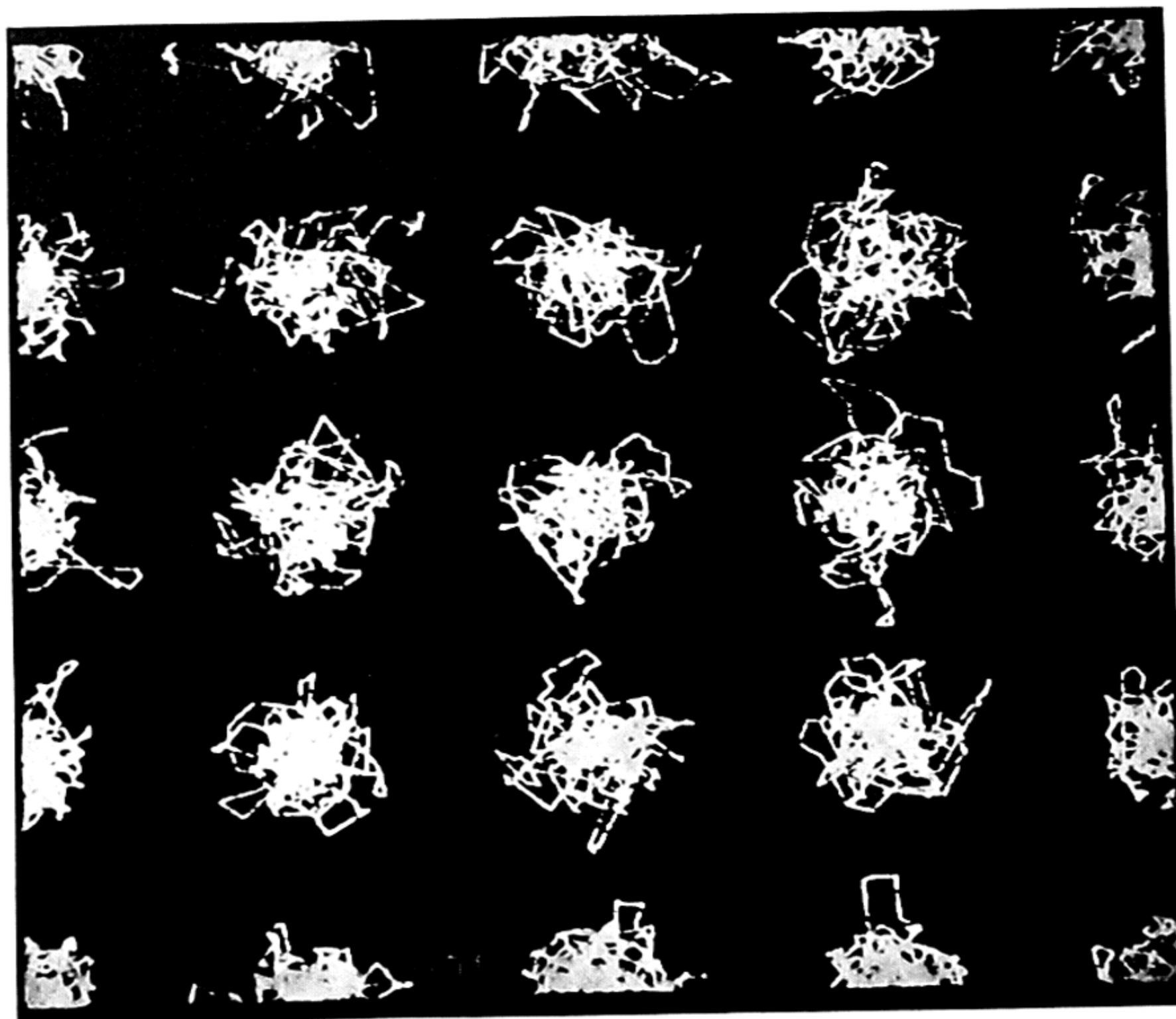
Monte Carlo calculations would of course be impossible to carry out by hand. Machines are quick and tireless, but they too have limitations. The largest existing computers can handle no more than about 500 particles in a reasonable calculating time. Machines soon to be available will have capacities for some 10,000. Even the latter figure is infinitesimal compared with the number of molecules in any weighable sample of matter. Yet it is remarkable how closely a system of just a few dozen particles can approximate the properties of real matter. One stratagem that helps make this

possible is the proper choice of "boundary conditions."

In any sizable piece of matter the overwhelming majority of molecules are inside, surrounded by other molecules and interacting with them. Only a tiny fraction is at the surface or boundary at any time. However, in our hypothetical system of a few dozen or a few hundred particles a substantial proportion will always be near the walls of their imaginary container. A random displacement is thus likely to bring them into contact with the boundary rather than with another particle. To avoid this atypical situation the computer is programmed to allow the particle to pass freely through the walls of the box. Whenever a particle moves out through one wall, it is immediately brought back in through the opposite wall. In effect we have put the opposite boundaries together, just as a string of beads is brought together by tying its ends. Now every particle is surrounded by neighbors and interacts



DIFFUSION OF PARTICLES in a specific time is plotted from molecular-dynamical calculation. Horizontal coordinate measures square of distance (in arbitrary units); vertical coordinate, numbers of particles that have moved various distances from starting point.



PATHS OF PARTICLES in molecular-dynamical calculation appear as bright lines on the face of a cathode-ray tube hooked to the computer. Each cluster in upper photograph represents two hard spheres, one behind the other. In solid state (*top*) particles can move only around well-defined positions; in fluid (*bottom*) they travel from one position to another.

with them instead of with the wall.

High-speed computers and the Monte Carlo method have enabled us to overcome many of the mathematical difficulties of statistical mechanics. Liquids and melting solids have been studied with the hard-sphere model and the pressure determined from the average distribution of pair separations.

Furthermore, it has been possible to consider more realistic intermolecular forces. For instance, two particles may be assumed to attract each other when they are closer than a certain distance, but repel each other when they actually collide. The shaking process now becomes more complicated. In the hard-sphere case one move is as good as another so long as it does not bring two particles into the same space. But when particles interact at a distance, some moves become more probable than others. Thus two particles within the range of attraction are more likely to approach each other than to separate. In carrying out the Monte Carlo calculations we assign probabilities reflecting the force that tends to encourage or inhibit each displacement: a given type of move might be allowed only every third time it came up.

With these refinements it is possible to take into account more of the properties of a real physical system, such as temperature. When attractive forces are considered, the likelihood of a particular change in the distribution of molecules depends on the temperature. Hence by varying the assigned probabilities we can play the Monte Carlo game at various temperatures. The distribution-of-distance plots can then be used to calculate the pressure of the system at different densities and temperatures, and the results compared with measurements on actual samples.

In this way the pressure of argon has been successfully calculated over a range of temperatures and densities that is wide enough to include the solid, liquid and gas phases. Moreover, the computations can easily be extended into regions of very high temperature and pressure, which are difficult to attain in the laboratory.

As has already been pointed out, the statistical mathematics of the Monte Carlo method produces accurate distributions of intermolecular distances. But it does not reproduce molecular motions. Hence it is restricted in its application to equilibrium properties such as pressure and energy. At the Lawrence Radiation Laboratory of the University of California we have undertaken a program to

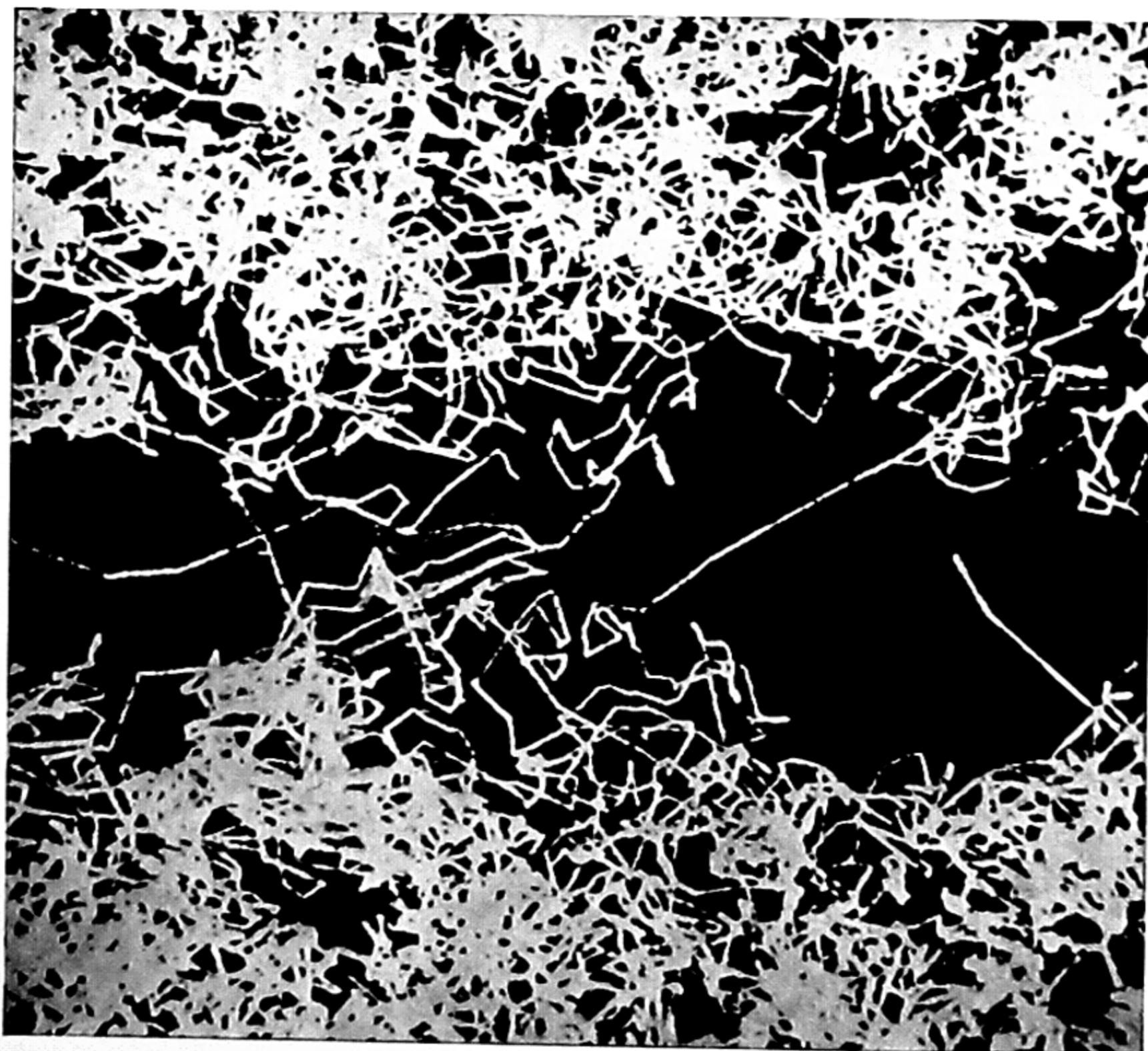
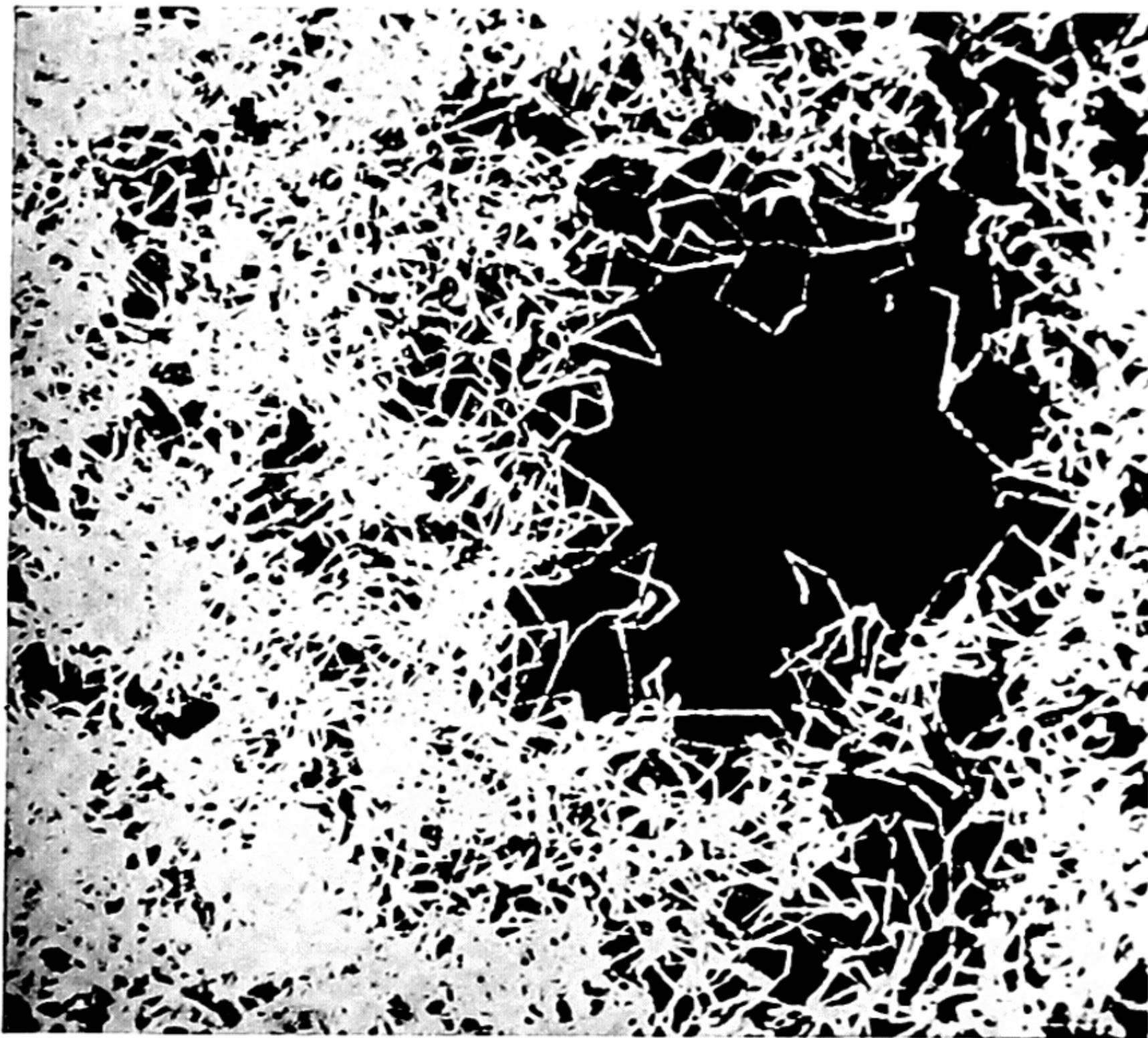
remove this deficiency. With the help of automatic computers we have been able to calculate the actual trajectories of a rather large number of individual particles. By means of such calculations it is possible to make theoretical determinations of properties that depend specifically upon details of molecular dynamics.

It turns out that a computer can follow the detailed motions of about as many particles as it can handle in the Monte Carlo method. We begin the calculation by assigning to each particle not only a position but also a velocity (*i.e.*, a speed and a direction of motion). Thereafter no element of chance is involved. The subsequent spatial patterns are predetermined. The machine calculates the path of every particle until two of them come close enough to exert a force on each other. Using Newton's laws of motion, the computer determines the effect of the forces on the two particles, and calculates their new velocities after the collision. Then it continues to compute all the trajectories until a second pair interacts, and so on. The same boundary conditions can be used as in the Monte Carlo method.

Of course we must make some assumption about the force between particles. The assumptions we make are not entirely realistic, but are designed to split a complex problem into simpler parts. Thus the mathematical complications arising out of realistic intermolecular forces can be isolated from those that are due simply to the large number of particles.

The two types of force that have been studied so far are the hard-sphere case and the so-called square-well interaction. In the latter there is no force between the particles until they approach to a certain distance. At this distance they abruptly attract each other. At shorter distances the attraction disappears, but if the particles then come close enough to touch each other they act like hard spheres and bounce apart. Since these particles move with constant velocity everywhere except when separated from one another by either the "attracting distance" or the "repelling distance," their motions are comparatively easy to compute. More realistic forces that vary smoothly with the distance of separation might be considered, but at the expense of more lengthy calculations.

The dynamical method is not restricted in its application to systems at equilibrium. It can, for example, be used to find how quickly the system reaches equilibrium. In one case a system of hard spheres was started out in the highly



LIQUID-GAS SEPARATION is illustrated by molecular-dynamical calculations on particles with square-well interaction. Dark area in upper photograph represents a gaseous bubble, surrounded by particles whose motions characterize a liquid. Lower photograph shows the system at a later time, when some particles have vaporized and passed through the bubble.

atypical condition in which all the particles had the same kinetic energy. The statistically expected distribution of energies was attained after very few collisions [see illustration on page 534].

Once the system reaches equilibrium, distance-of-separation snapshots can be taken just as in the Monte Carlo method. These snapshots yield the same information as those achieved by statistical techniques. For example, they might be used to calculate the pressure. The dynamical method furnishes an additional way of arriving at the answer. If the system is considered to have reflecting walls, the pressure can also be calculated by considering the impacts of the molecules on the walls. Pressures in hard-sphere systems have been determined by both the Monte Carlo and the dynamical methods with very close agreement.

Because the configurations calculated by the dynamical method are in the correct temporal order, they enable us to study the time-dependent behavior of the system. This is important even after equilibrium has been attained. One example of time-dependent behavior is diffusion. At equilibrium the average total force on a molecule is zero, because the forces exerted on it by its neighbors are as likely to come from one side as another. Nevertheless any specific particle is likely to experience a succession of forces that leave it, after an interval of time, with a net displacement from its original position. In other words, it will diffuse. By noting the positions of the particles in successive snapshots, we can establish their rate of diffusion.

Albert Einstein's theory of diffusion predicts that the average of the squares

of the net displacements of the particles is proportional to the time during which diffusion takes place. Molecular-dynamical studies have shown that the theory is true for hard-sphere fluids and have also determined the constant of proportionality for a wide range of fluid densities.

So far the most extensive study carried out by the dynamical method has been the calculation of the pressure of a hard-sphere system over a large range of densities. The main purpose was to shed some light on a long-debated question: Can a hard-sphere material have a sharp freezing point? Does the change from the disorganized configuration of the molecules in a liquid to the ordered lattice of a crystal occur gradually or suddenly, if the molecules are assumed to have no attractive forces?

If hard spheres are in an orderly arrangement at a particular density, the system will have a lower pressure than if they are arranged at random, as in a liquid. Systems with as few as eight and as many as 256 hard-sphere particles have been studied. All exhibit sharp jumps in pressure when the density is held fixed somewhere between that of a crystal and that of a liquid [see illustration on page 535]. The pressure-jumps are associated with abrupt changes in the system from the orderly, crystalline phase to the random, fluid phase and vice versa. The size of the jump depends on the number of particles; the smaller the number, the larger the jump.

In all cases the system changed as a whole from one phase to the other: it was either entirely crystalline or entirely fluid. If a larger number of particles was used, we might expect to see a two-phase equilibrium, that is, part of the

system crystalline and part of the system fluid. This would correspond to the physical situation in which ice and water, for example, can exist side by side in equilibrium at the freezing point.

In order to demonstrate the phase changes more vividly, a display system similar to a television picture-tube was hooked up to the computer. Each particle in the system was then represented by a dot on the face of the tube. By focusing a camera on the screen and leaving its shutter open, it was possible to record the trajectories of the moving dots on film. The photographs on page 536 show the imaginary material in the crystalline and the fluid phases.

If we want to study the liquid-gas phase transition, we must consider molecules with attractive forces. Such a study has been made with 32-particle systems with square-well interactions [see photographs on page 537].

We have mentioned only some of the molecular-dynamical calculations that have been carried out thus far. They represent an ideal example of the application of automatic computers to scientific research. A knowledge of the physical laws is assumed. The straightforward but tedious mathematical calculations are carried out by machines that are specifically designed for that purpose.

Perhaps it is not too much to expect that the information obtained by means of computing machines may play a role analogous to that of laboratory experiments in the development of theory. When we have built up a sufficiently large body of numerical computations, we may be able to discern generalizations that are not apparent to us now.

The Authors

B. J. ALDER and THOMAS E. WAINWRIGHT are both theoretical physicists on the staff of the Lawrence Radiation Laboratory in Livermore, Calif. Alder was born in 1925 in Duisberg, Germany. He was a Swiss citizen, and received his high-school education in Switzerland. He came to the U. S. in 1941, and from 1944 to 1946 he was an electronic technician in the Navy. In 1947 he acquired a B.S. in chemistry from the University of California. He took his M.S. there in 1948 in chemical engineering, and shifted to the California Institute of Technology for his Ph.D. in 1951. He was an instructor in chemistry at the University of California from 1951 to 1954, held a one-year Guggenheim Fellowship at the universities of Cambridge and Leiden, and in 1955 took up his present work at the Lawrence Radiation Laboratory. Wainwright is 32 years old. He received his college education at the University of Utah and Montana State College, and took his Ph.D. in physics at the University of Notre Dame in 1953. He joined the Lawrence Laboratory in 1954.

Bibliography

- EQUATION OF STATE CALCULATIONS BY FAST COMPUTING MACHINES. Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller and Edward Teller in *The Journal of Chemical Physics*, Vol. 21, No. 6, pages 1,087-1,092; June, 1953.
- FURTHER RESULTS ON MONTE CARLO EQUATIONS OF STATE. Marshall N. Rosenbluth and Arianna W. Rosenbluth in *The Journal of Chemical Physics*, Vol. 22, No. 5, pages 881-884; May, 1954.
- MONTE CARLO EQUATION OF STATE OF MOLECULES INTERACTING WITH THE LENNARD-JONES POTENTIAL. I. A SUPERCRITICAL ISOTHERM AT ABOUT TWICE THE CRITICAL TEMPERATURE. W. W. Wood and F. R. Parker in *The Journal of Chemical Physics*, Vol. 27, No. 3, pages 720-733; September, 1957.
- RADIAL DISTRIBUTION FUNCTION CALCULATED BY THE MONTE CARLO METHOD FOR A HARD SPHERE FLUID. B. J. Alder, S. P. Frankel and V. A. Lewinson in *The Journal of Chemical Physics*, Vol. 23, No. 3, pages 417-419; March, 1955.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

DATE LOANED

Acc. No. _____

[illegible]

THE FORCE BETWEEN MOLECULES

by Boris V. Derjaguin

Its nature has long been a mystery. Physicists in the U.S.S.R. have now succeeded both in measuring the intermolecular force and in demonstrating that it is purely electromagnetic in origin.

One of the great traditions of science has been the study of the forces that hold the parts of the material world together. The tradition runs from Isaac Newton's work on gravity and celestial mechanics to the modern examination of the nucleus of the atom. In this long line of inquiry there has been, until very recently, an odd gap. The smallest units of matter, in its everyday forms, are molecules; and the force between molecules determines many characteristics of the substances with which we are familiar. Yet physicists have been unable either to measure intermolecular forces experimentally or to deduce them from theory.

These forces are not the chemical attractions, or valence bonds, that link atoms in compounds. They are forces of longer range that draw together the molecules of solids and liquids, and, to a much smaller extent, of gases. Valence forces can be "saturated"; when the available bonds of an atom are fully satisfied by attachment to its partners in a compound molecule, it exerts no force on other neighboring particles. Thus the oxygen atoms in two molecules of water do not attract each other. On the other hand, every molecule in a sample of water exerts some attraction on every other molecule.

Without these long-range forces there would be no sharp boundary between liquids and their vapors. Intermolecular forces are responsible for surface tension, capillary action, adsorption and

other surface phenomena. They determine most of the properties of liquids: viscosity, heat of evaporation, solubility in other liquids. They cause colloids to coagulate. Each of these phenomena has been intensively studied for years. But without an understanding of their underlying principle there could be no fundamental and unifying theory.

In the U.S.S.R. during the past few years there have been two significant break-throughs. Irene Abrikosova and the author have succeeded in measuring the molecular force directly. Following this development the theoretical physicist E. M. Lifshitz derived a mathematical formula for the force of molecular attraction from very general principles.

To put these developments in proper perspective let us begin with some history. The earliest speculations about molecular forces were those of the 18th-century French mathematicians Alexis Claude Clairaut and Pierre Simon de Laplace. Theirs was the triumphant era when Newton's law of universal gravitation seemed capable of explaining the motions of heavenly bodies down to the minutest detail. By analogy with gravity Clairaut and Laplace assumed the existence of an attractive force acting along the line between the centers of molecules. It was apparent, however, that such a force must fall off not as the square of the distance of separation, as in the case of gravity, but more sharply. Furthermore, the constant of propor-

tionality could be different for different molecules.

In arithmetical terms Newton's law contains the fraction G/r^2 , where G is the constant of proportionality. What makes gravity comparatively easy to deal with is that G is a universal constant whose value is determined for all masses simply by measuring the force between any two known masses at any known distance (r). On the other hand, the assumed law of molecular force has the form C/r^n , where both C and n are unknown, and where C presumably differs for different kinds of molecules. It is of course impossible to pick out two individual molecules, hold them at a known distance and measure the force between them. Hence the unknowns cannot be directly determined. Nor, it turns out, can they be determined indirectly through bulk properties that depend on molecular forces. Measurements of surface tension and similar quantities cannot even settle the power law, let alone establish the constant C . No real progress could be made without a better theoretical platform from which to attack the problem.

At the end of the last century it seemed that the classical theory of electromagnetism could provide such a platform. The laws governing electric and magnetic fields and their interactions with material bodies had been worked out. Molecules were known to be composed of electrically charged particles. The visible radiation given out by

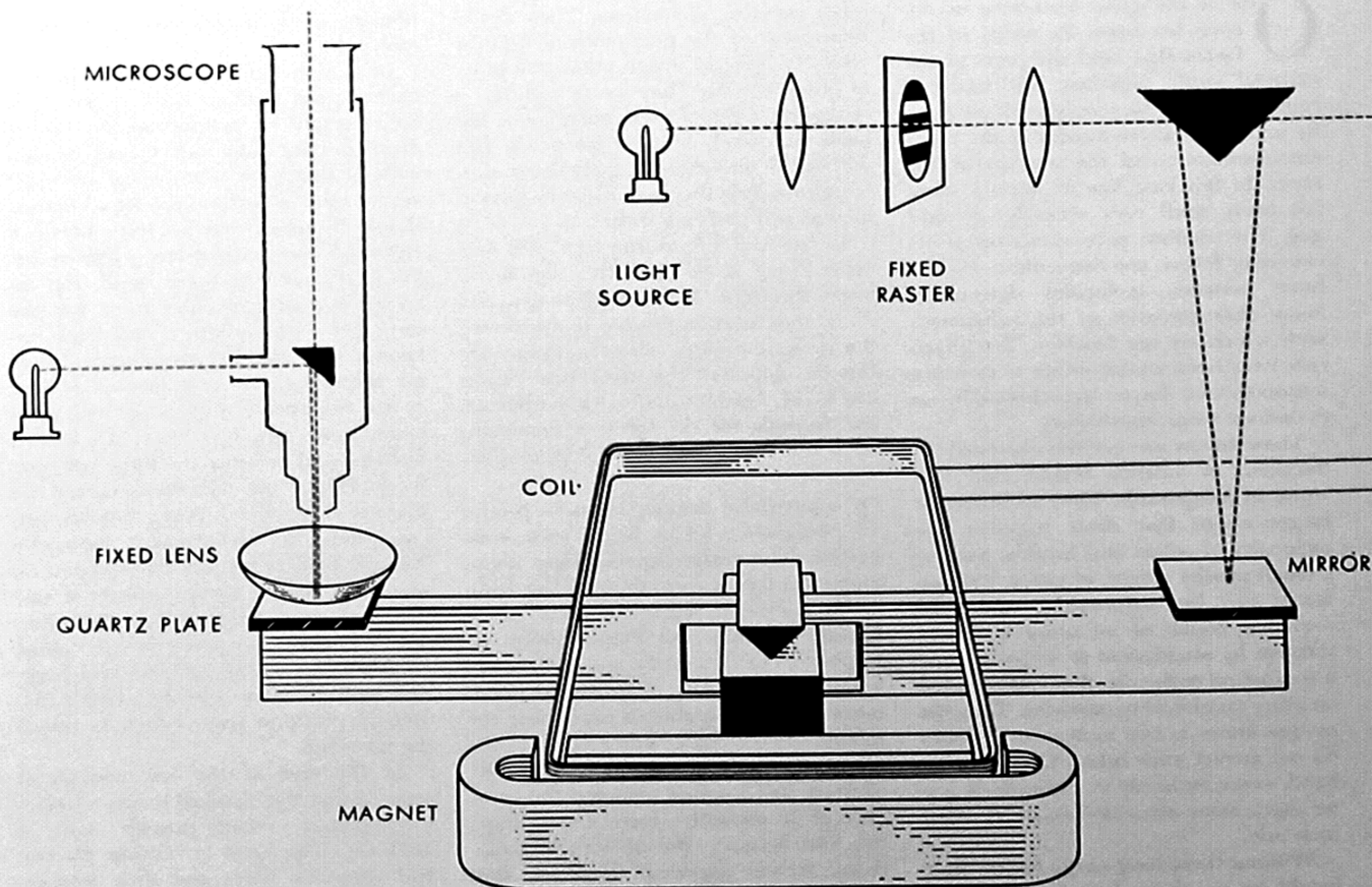
molecules had just been proved to be electromagnetic waves. Perhaps electromagnetism could account completely for the intermolecular force. (It should be realized that this was just an assumption. It was also possible that some new force, previously unknown, might come into play at the molecular level, as has since proved to be the case at the nuclear level.)

One of the first workers to see clearly the implications of electromagnetic theory was the Russian physicist P. N. Lebedev, best known as the first experimenter to measure the pressure exerted by light. In 1894 he wrote, with prophetic insight: "Hertz's research into the interpretation of light waves as electromagnetic processes conceals another problem, as yet untouched. This is the problem of a radiation source, of those processes which a molecular vibrator undergoes when it emits light energy to its surroundings. On the one hand this problem leads into the field of spectro-

scopic analysis. On the other hand, almost entirely unexpectedly, it leads to one of the most complicated questions of modern physics, the science of molecular forces. This situation results from the following concepts. From the point of view of the electromagnetic theory of light one may state that two radiating molecules are two vibrators in which electromagnetic vibrations are excited. They must therefore experience forces resulting from the electromagnetic interactions of the varying electrical currents (according to Ampère's law) and varying charges (according to Coulomb's law) in them. We may therefore also assert that there must then exist molecular forces whose origin is directly related to radiative processes. . . . The most interesting as well as the most complicated case is that of a physical body in which many molecules act simultaneously on one another and are so closely packed that their vibrations are not independent. If it ever becomes possible to solve this problem completely, we will be able

to use spectroscopic data to predict the magnitude of the molecular forces due to the mutual radiation of the molecules, to calculate the temperature dependence of these forces, and by comparing these calculated quantities with experiment to solve the problem at the root of molecular physics. This problem is whether all so-called molecular forces reduce to the already known and above-mentioned action of radiation—to electromagnetic forces—or whether they involve other forces whose source is still unknown."

Lebedev came as close to the latest views on the nature of molecular forces as was possible before the development of quantum mechanics. But his statement, while ahead of its time, was very far from a quantitative law. As so often happens in science, the next step was backward. The first quantitative theories of molecular forces treated them as entirely electrostatic rather than electromagnetic. These static theories began to be developed shortly after



CONTROLLED BALANCE for measuring the molecular attraction between macroscopic bodies is illustrated schematically. The broken colored line at left is the path of the light beam used to de-

termine the width of the gap between the bodies. The broken colored line at top right is the path of light beam by which the position of the balance arm is set and automatically stabilized. Solid

Ernest Rutherford, by his discovery of the nucleus, showed how electric charge is distributed in the atom. However, they did not reach their final form until 1930, when the German physicist Fritz London applied to them the newly discovered principles of quantum mechanics.

According to London's theory, the force between two molecules is C/r^7 . The force varies inversely as the seventh power of the distance between their centers. The constant C depends on certain electrical properties of the molecules, including their "polarizability." This measures the degree to which an electric field distorts a molecule, shifting its electrons with respect to the positive nuclei of its atoms.

Unfortunately the polarizabilities of individual molecules are themselves unknown quantities, so we cannot use London's formula to calculate the absolute value of the force. If it were possible to compare molecular forces at various distances, the seventh-power law could at

least be tested, but for a long time there was no way of doing this either. Attempts were made to verify the formula indirectly, for example by measuring the amount of heat energy required to vaporize a liquid (*i.e.*, pull its molecules apart). However, it can be proved that most of the energy is used up in the interval when the molecules are still within a diameter or two of each other. At such close distances London's theory is not strictly applicable, and even the concept of distance between molecules (which are not really spheres) is not clearly defined.

Despite the lack of experimental confirmation, or perhaps because of it, the London theory was generally accepted for almost 20 years. Yet it was clearly unrealistic for large distances of separation as well as for very small ones. Being an electrostatic theory, it regarded force as communicated instantaneously between molecules. In fact electric forces travel at the finite speed of light (186,000 miles per second).

Now the "large" distances just mentioned are large only compared with molecular diameters. On the everyday scale of distance the attractive force between molecules falls away to essentially nothing at a very small separation indeed—a few ten thousandths of a millimeter. The time required by an electromagnetic impulse to travel such a distance is of the order of a ten-million billionth of a second. How can so tiny an interval make a difference in the force?

To understand this let us recall Lebedev's picture of "molecular vibrators" sending out radiation. The molecules are like radio antennas in which oscillating electric charges emit a train of electric and magnetic vibrations. When the vibrations from one antenna reach a second, they set its charges into oscillation. These oscillations become in turn a source of waves that reach back to the first antenna, exerting a force on its moving charges. Thus the two antennas interact by exchanging radiation. The strength of the interaction varies with the relative phase of the arriving waves and the oscillating charges; the interaction is stronger if the waves and particles swing back and forth together rather than being totally or partially out of step. But the relative phase depends on the distance between antennas and on the frequency of the vibrations; in other words, on the number of wavelengths in the distance of separation. If, for example, the antennas are half a wavelength apart, the emitted and absorbed waves will be exactly out of phase. At smaller

(or larger) fractions of a wavelength the shift will be correspondingly less. At very small fractions of a wavelength the phase shift is negligible.

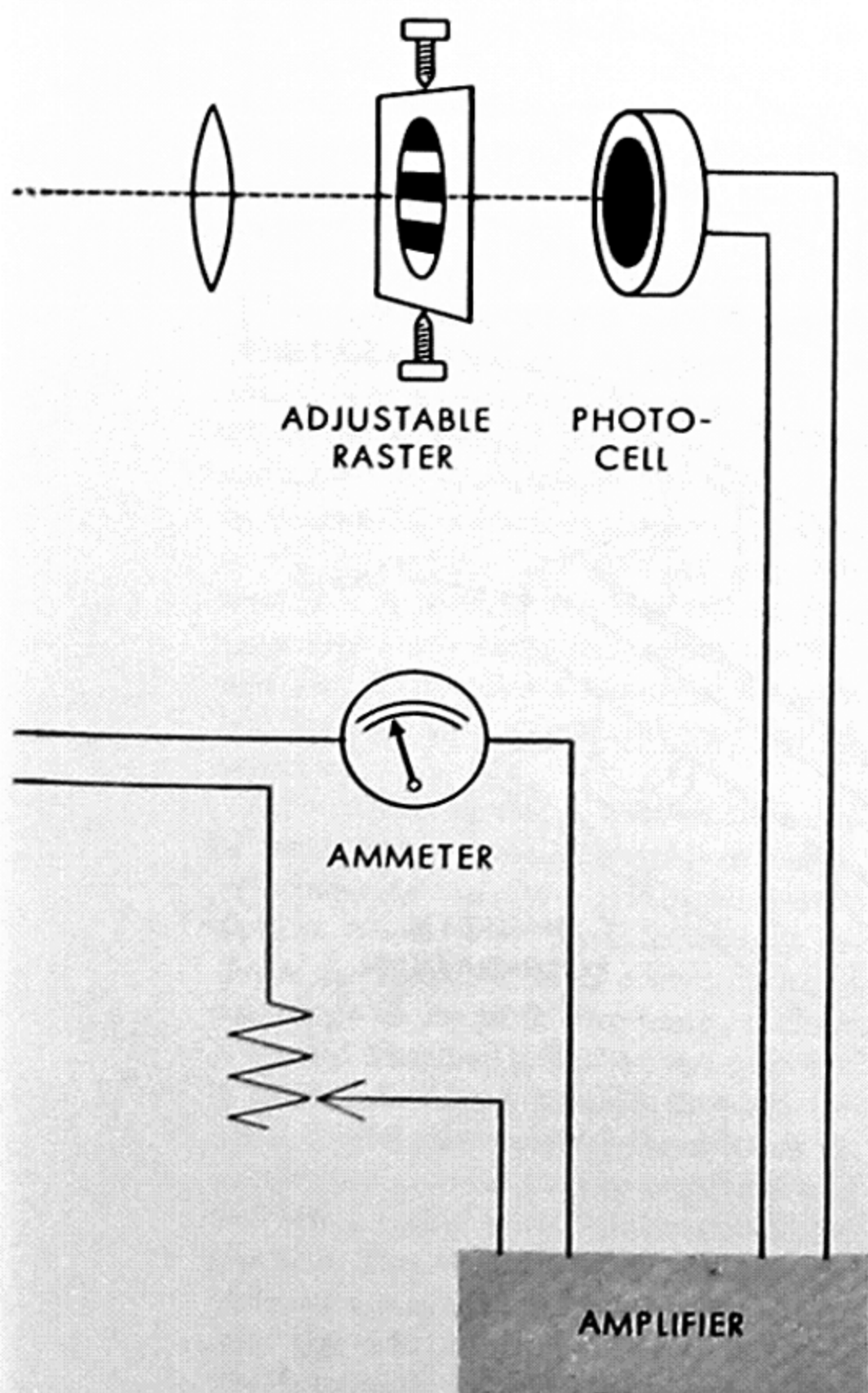
The wavelength of most radio waves is measured in meters or centimeters. But the radiation sent out and absorbed by molecules are light waves, only a ten thousandth of a millimeter long. Thus if two molecules are separated by a ten thousandth or even a hundred thousandth of a millimeter, there will be considerable phase shift in the exchanged radiation. The shift must be considered in computing the force.

The first theoreticians to take the phase shift into account in their calculations were the Dutch physicists H. B. G. Casimir and D. Polder, who worked out an electromagnetic theory of molecular force in 1948. In their calculations they did not, of course, use the classical picture of radiation that we have just outlined but rather the ideas of quantum electrodynamics. These differ from the older view in two important respects. First, molecules do not radiate (or absorb) continuously, but in discrete amounts. Moreover, the waves they send out travel in packets or photons. Second, molecules can interact electrically without actually emitting or absorbing energy. In that case they are said to exchange "virtual" photons.

At relatively large distances the Casimir-Polder electromagnetic theory gives the formula K/r^8 for the attractive force between two molecules. Thus when the effects of phase shift are allowed for, the force decreases with the eighth power of the distance rather than with the seventh, as in London's formula. The constant K is different from London's C , but it also contains the polarizability of the molecules. Therefore it cannot be evaluated directly either.

We shall see shortly that each of the two formulas appears to be correct for an appropriate range of the distance r . London's formula applies when the distance is small enough to make the phase shift of the exchanged radiation negligible. The Casimir-Polder result describes the force at larger distances, where the phase shift cannot be disregarded.

Both theories, however, concern only pairs of isolated molecules. They do not themselves give the force of molecular attraction between two condensed bodies, each containing many molecules packed close together. Yet this is the only force one can hope to observe directly, for example by measuring the at-



colored lines trace the electrical portion of the control circuit. The operation of the system is described in detail in the text.

traction between two solids separated by an extremely narrow gap. (This is precisely the measurement that my colleagues and I have finally succeeded in making.) Furthermore, it is this force that presumably causes the tiny particles of a colloid (which are nevertheless large enough to contain many molecules) to stick together and coagulate. Thus it is fundamental to the theory of the colloidal state.

If we assume that molecular forces are not influenced by interactions with neighbors, that each molecule in one dense body exerts the same force on each molecule of a nearby body as if they were so many separate pairs, then it should be possible to obtain the total force simply by addition. Some workers have made this assumption and have proceeded to compute the total force between condensed bodies both by the London and the Casimir-Polder formulas. However, there is good reason to

believe that the forces are not strictly additive, so these computations cannot give the right answer.

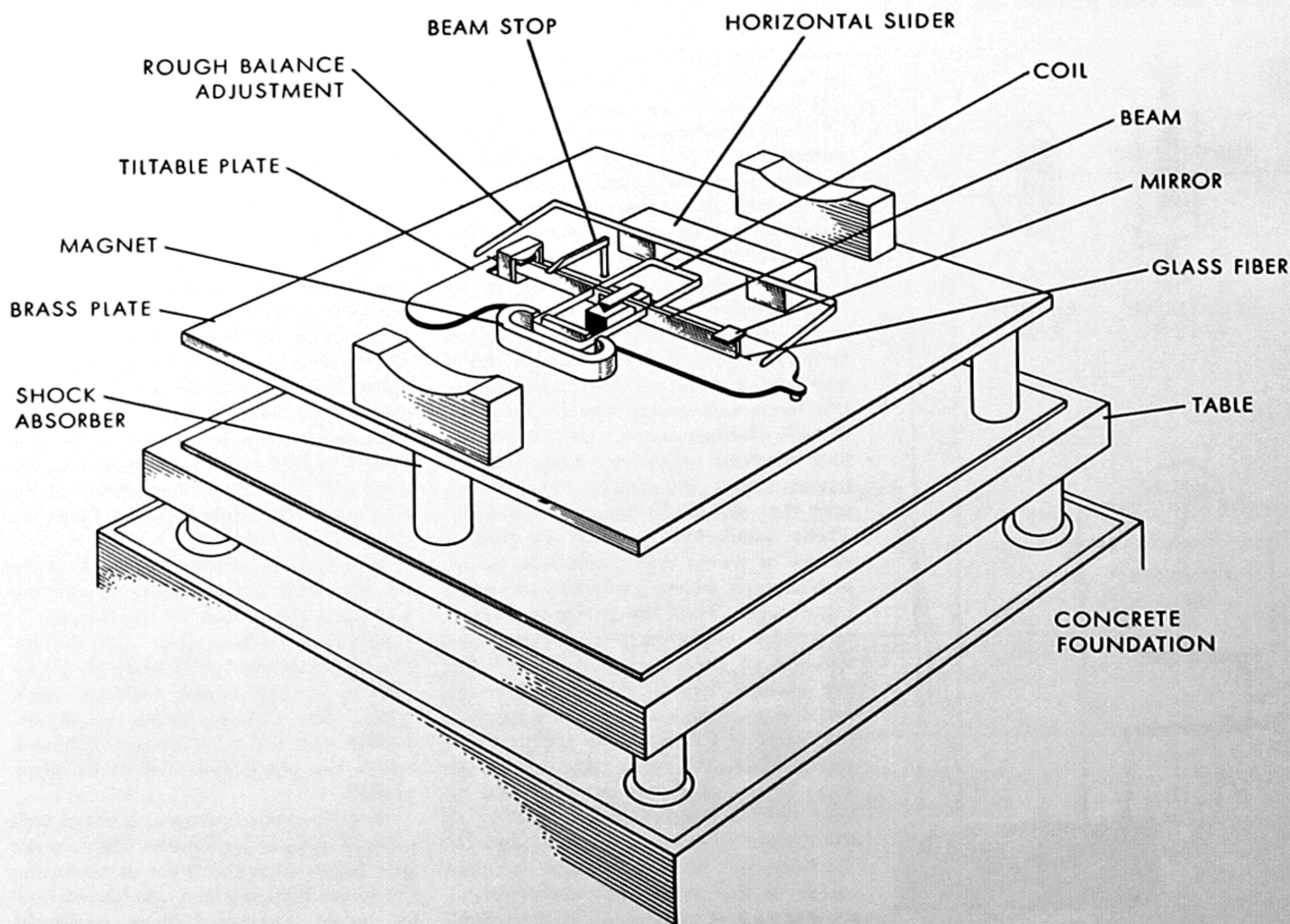
Clearly the individual-particle approach to molecular forces had proved unprofitable. Its formulas contained unknown quantities, and so could not be evaluated. Even if they could, there was no way of proceeding from isolated pairs of molecules to the gross samples of matter with which experimenters are obliged to deal.

This was the unsatisfactory state of theoretical knowledge when Mrs. Abrikossova and I set out in 1951 to measure the molecular attraction between a pair of solid objects. With the help of Fanny Leib, a colleague at the Academy of Sciences in Moscow, we were able to design an apparatus that met the stringent, and in some ways contradictory, requirements of the job.

In the first place we were trying to

measure a small force—as little as a ten thousandth of a gram. This in itself is not too difficult; microchemists routinely deal with far smaller forces. But in our problem the attraction to be measured shows up only when the two bodies are extremely close together—within a few ten thousandths of a millimeter, or little more than a thousandth the thickness of a human hair. Furthermore, the force varies enormously as the distance of separation is changed. It is precisely this variation that we wanted to determine. Thus we needed a way to set the width of the gap accurately at any desired distance, and to maintain the separation against the attraction force being measured. Not only that; the equilibrium had to be extremely stable so that chance variations in the gap were corrected before they could build up.

The requirements of sensitivity, stability and rapid response are not easy to reconcile. A sensitive balance tends to



BALANCE IS MOUNTED on a brass plate connected through shock absorbers to a table, which in turn rests on a concrete foundation. Tilttable plate allows the entire balance to be rotated,

changing the position of the gap with respect to the surfaces of the objects whose attraction is being measured. Glass fiber extending through the balance arm can be shifted for rough balancing.

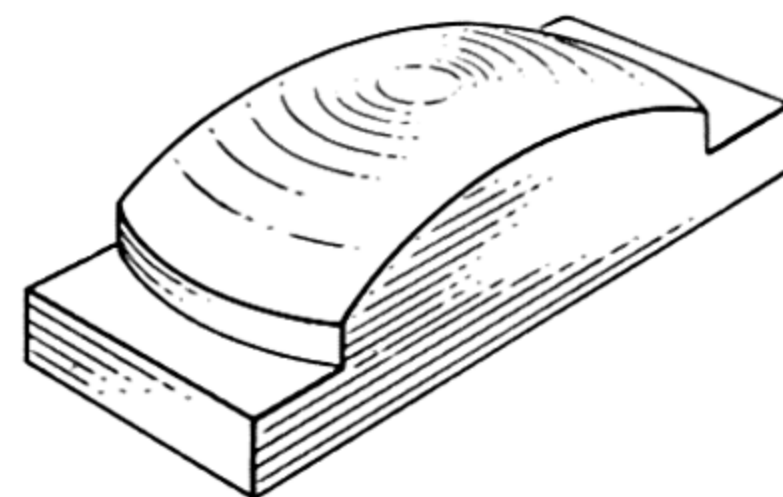
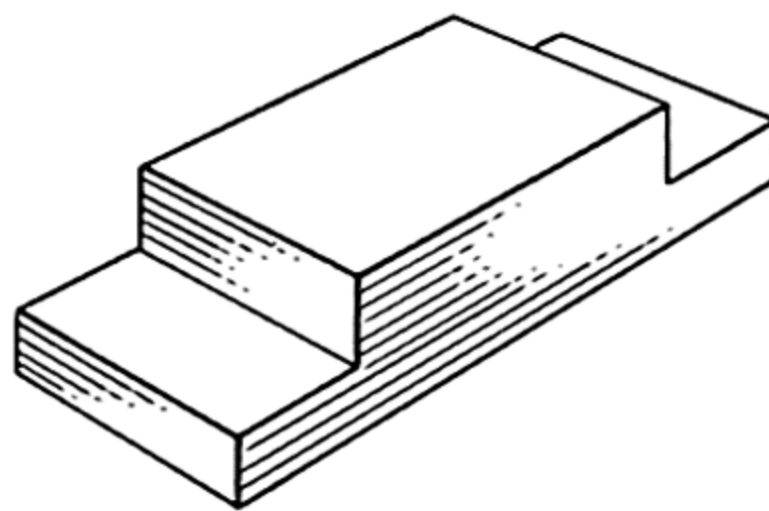
swing widely in response to a small change in force. Our apparatus achieved these diverse aims by means of automatic control.

The bodies whose mutual attraction we measured were a flat plate and a plate with a spherical surface. (It is easier to set the distance between a sphere and a plane than between two planes, which must be held parallel.) The flat plate was fixed to one end of a balance arm three centimeters long, weighing about a tenth of a gram and balanced on a fulcrum made of a wedge-shaped agate bearing. Above the flat plate was mounted the rounded plate in an adjustable bracket [see illustration on pages 542 and 543.]

To determine the width of the gap between the two surfaces we made use of the optical phenomenon known as Newton's rings. When light of a single color shines through a lens-shaped disk held near a flat surface, interference between the rays reflected from different surfaces produces a pattern of concentric light and dark circles [see top illustration on next page]. The diameter of the circles depends on the distance between the surfaces. We observed the Newton's rings in our sample through a microscope, accurately determining their diameter and thus measuring the gap width.

The automatic-control circuit, which is the heart of the arrangement, accomplishes a quadruple purpose: (1) it allows us to set the gap at any desired distance, (2) it maintains the separation by producing a force equal and opposite to the attraction, (3) through negative feedback it corrects any drift in the balance arm away from the preset position and (4) it provides a means of measuring the opposing force, and thus the attraction.

The circuit works as follows. A beam of light shines through a grid, or raster, of alternate opaque and transparent stripes. A mirror on the balance arm reflects the beam through a lens, forming an image of the grid. The image falls on a second identical grid, beyond which is a photocell. Light passing through the second grid and reaching the cell sets up an electric current that is amplified and fed into a coil of wire rigidly attached to the arm. This coil is pivoted between the poles of a magnet and thus tends to turn, like the coil in an ammeter, carrying the arm with it. The turning force depends on the amount of current passing through the coil, and is arranged to be opposite in direction to the turning force



TEST BODIES for the measurement of molecular force were a highly polished flat plate (*left*) and an equally smooth spherical lens (*right*). In most trials the bodies were made of quartz, but some other materials were tested. The lens diameter was also varied.

on the arm due to molecular attraction.

The current in the coil is determined by the amount of light that reaches the photocell. Let us see what controls this. If the image of the first grid falls on the second grid in such a way that the transparent stripes of the image exactly cover the black stripes of the grid, no light gets through. The current through the coil is zero. Rotating the arm slightly shifts the angle of the mirror, thus moving the image. Now part of the transparent stripes in the image overlap the transparent stripes on the second grid, and some light passes through. The farther the arm turns, the more light passes, and the stronger the current.

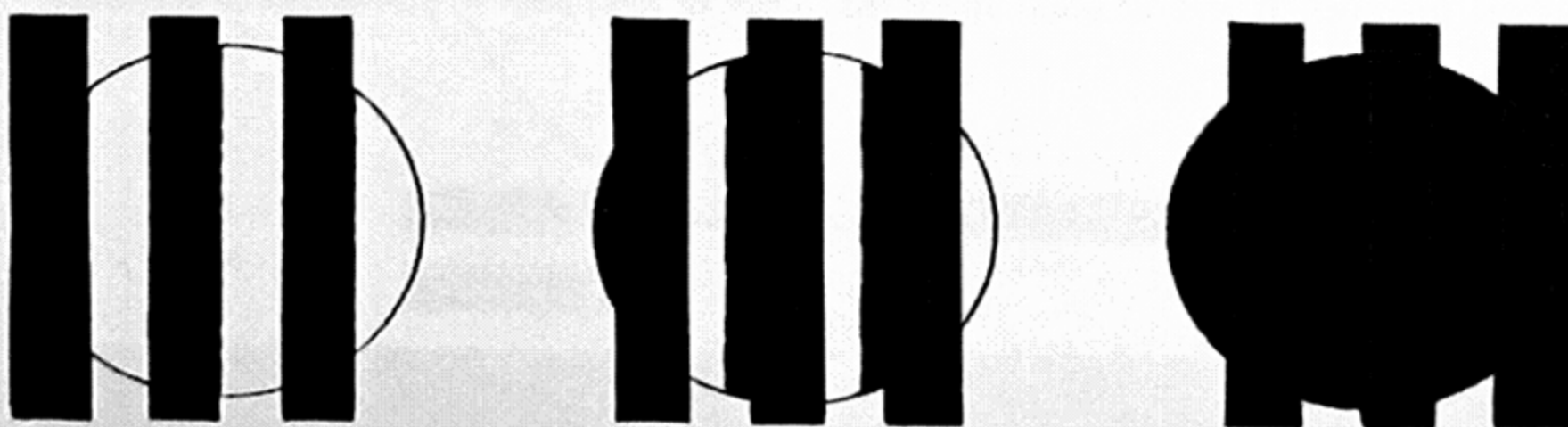
In setting the gap initially at the desired width, we shift one grid by means of a micrometer screw forcing the arm to turn as the equilibrium position varies. We then read the current on an ammeter, thereby determining the force at this gap width. If a stray vibration alters the distance of separation, the current changes in such a way as to bring the arm back to its initial position. Without the control circuit, the balance arm takes several seconds to swing back and forth when it is displaced from equi-

librium. When the control is switched on, this period is reduced to a thousandth of a second.

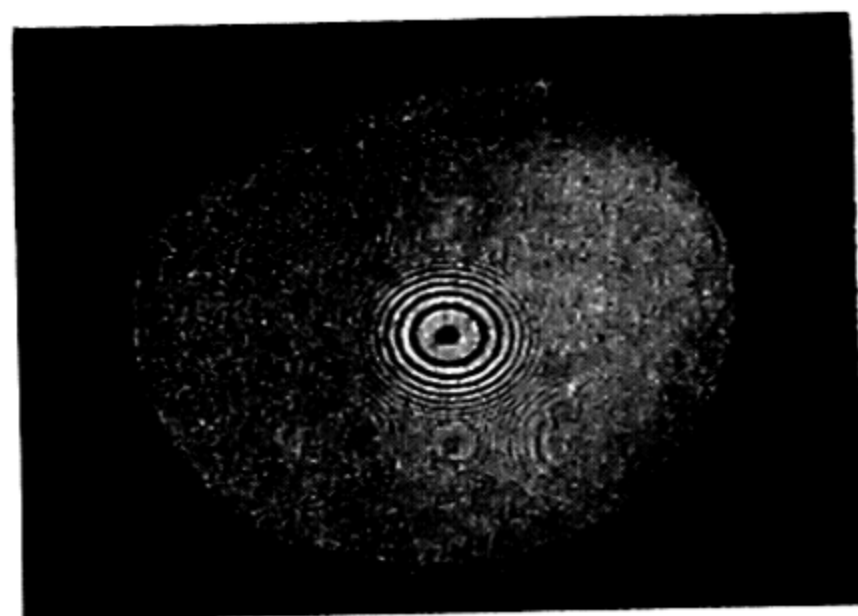
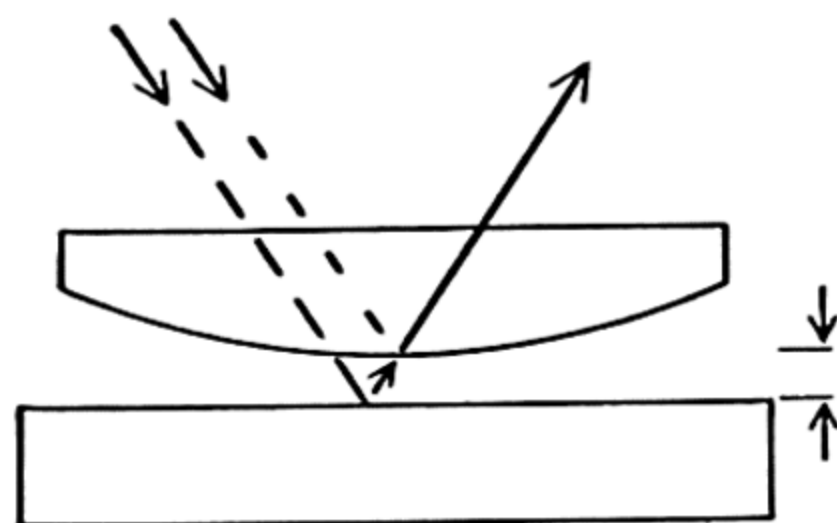
Fluctuations of the current in response to outside vibrations set the eventual limit on the accuracy of our measurement. To minimize the disturbances we mounted the balance on a heavy pedestal, connected by means of a hydraulic shock-absorber to a cement pier sunk in the ground. Air currents were eliminated by enclosing the apparatus in an evacuated chamber. This also reduced the viscous drag of the air in the narrow gap on the motions of the balance arm.

The most annoying and persistent difficulty we encountered in getting the balance to work properly was in trying to keep the surfaces of the plates free from both electrostatic charge and dust particles. Electrostatic forces are thousands of times stronger than the molecular attraction, and would completely obscure it. Dust particles on the surfaces would change the effective gap width.

To eliminate dust we cleaned the plates with cotton dipped in ether. Charge was removed by ionizing the air in the gap with a radioactive material. Repeated treatments were often re-



OVERLAPPING GRIDS, or rasters, determine the amount of light reaching the photocell in the balance-control circuit. Image of opaque bands (*colored stripes*) on one grid falls on second grid (*circle*). Light transmitted varies from a maximum when the two sets of bands coincide (*left*) to zero when image bands wholly fill transparent spaces (*right*).



NEWTON'S RINGS (*photograph at right*), used to determine gap width in the author's experiment, are formed when monochromatic light passes through a lens and is reflected from a flat plate. Some light is also reflected from lens surface, and the rays at any given point reinforce or interfere depending on the width of the gap between lens and plate.

quired, because the surfaces tend to become charged as they are cleaned, and to get dusty when the charge is removed.

In our first experiments in 1951 we measured forces that were about 5,000 times as large as predicted by any theory. Experimenters in the Netherlands got the same sort of result at this time. When we finally were able to get rid of all electric charge, however, the spurious forces disappeared. Other workers have duplicated our results.

We measured the force between two quartz plates, between a quartz and a chromium plate and between crystals made of a mixture of thallium bromide and thallium iodide. For gaps ranging from about two to four ten-thousandths of a millimeter the force varied from approximately 20 to two ten-thousandths of a gram [*see graphs on opposite page*]. The force obviously depended on the types of molecule of which the plates were composed.

When we obtained our first successful measurements, we could compare them only with the London and Casimir-Polder theories. It was at once apparent

that the Casimir-Polder formula fit our results better than did London's. (A gap of a ten thousandth of a millimeter is, of course, large on the scale of molecular diameters.) However, we could not check the formula absolutely, for the reasons already mentioned: it contained undetermined constants, and there was no rule for applying the microscopic equation to macroscopic bodies.

When Lifshitz heard of our results, he was encouraged to attack the theoretical problem anew. In 1955 he developed a remarkable method for finding the molecular interaction of two macroscopic bodies. The method ignores the individual particles completely, and depends only on macroscopic properties, which can be measured experimentally. We can do no more here than give a hint as to the basis of his highly abstract, mathematical approach.

As every amateur radio-operator knows, the effort to receive extremely weak signals from distant stations is in the end frustrated by noise originating in the receiver itself. This noise arises out of the thermal vibrations of the mole-

cules in the wiring and other parts of the set. The resulting electromagnetic fluctuations are noticeable when a radio receiver emits a hiss. But they exist silently in any material body. The only way to suppress them completely is by cooling the bodies to absolute zero and thus removing all their thermal energy.

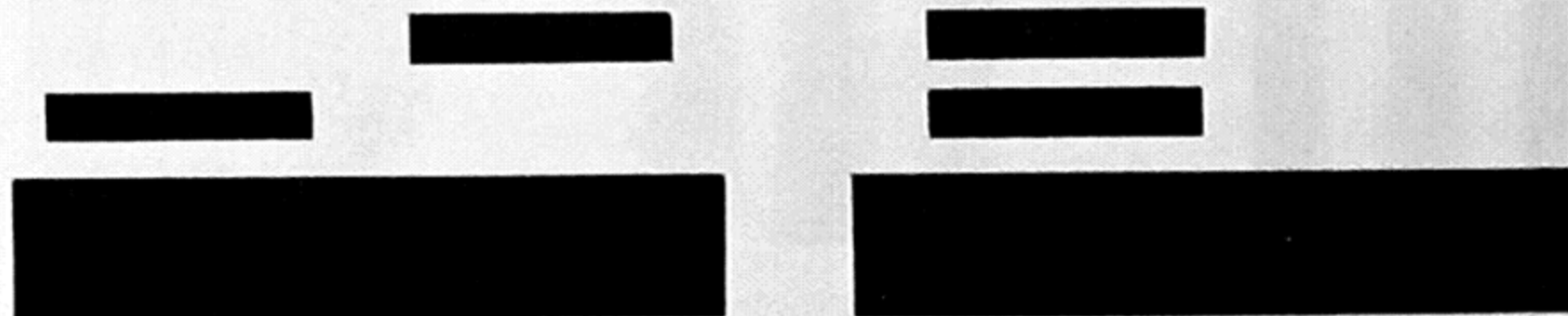
According to quantum theory, however, there must be fluctuations in the electromagnetic field even at absolute zero. These fluctuations are the result of the so-called zero-point energy of the electrons. By definition this energy cannot be detected or extracted from matter; it is represented entirely by the emission and absorption of virtual photons. Nevertheless it is there, and it is a means of interaction of the particles.

Lifshitz realized that such universal zero-point vibrations should in fact account for the molecular force. In his theory, however, he does not deal with discrete particles and light quanta, or photons, but with continuous matter and fields, along the lines of classical electromagnetic theory. He considers two closely spaced bodies and calculates the electromagnetic fields produced in the narrow gap between them and in the space around them by the fluctuations in the various regions of the material. From the difference between the field strength in the gap and in the surrounding space the force can be calculated.

The formulas are very complicated, but they do not contain any quantities that cannot be measured. To find the force it is necessary to know only the wavelengths absorbed by the materials (*i.e.*, their absorption spectra in the infrared, visible and ultraviolet regions) and their polarizabilities or "dielectric permeabilities." Unlike the polarizability of individual molecules the dielectric permeability of a macroscopic sample of a substance is easily determined.

Not only does the theory account for a force at absolute zero; it also shows that thermal oscillations contribute very little to the force at higher temperatures. In the range near absolute zero the force is almost independent of temperature. As the temperature is increased, the force does vary, but only through an indirect effect. Changing the temperature changes the quantum state of the electrons in the bodies, altering their absorption spectra and hence the force.

It is surely remarkable that a purely molecular effect can be calculated by ignoring the very existence of particles. (They are ignored except for one point:



NONADDITIVITY of molecular force on the macroscopic scale is demonstrated by the fact that the total force of attraction between the large block and the small ones is changed by shifting the small blocks from the arrangement shown at left to the one shown at right.

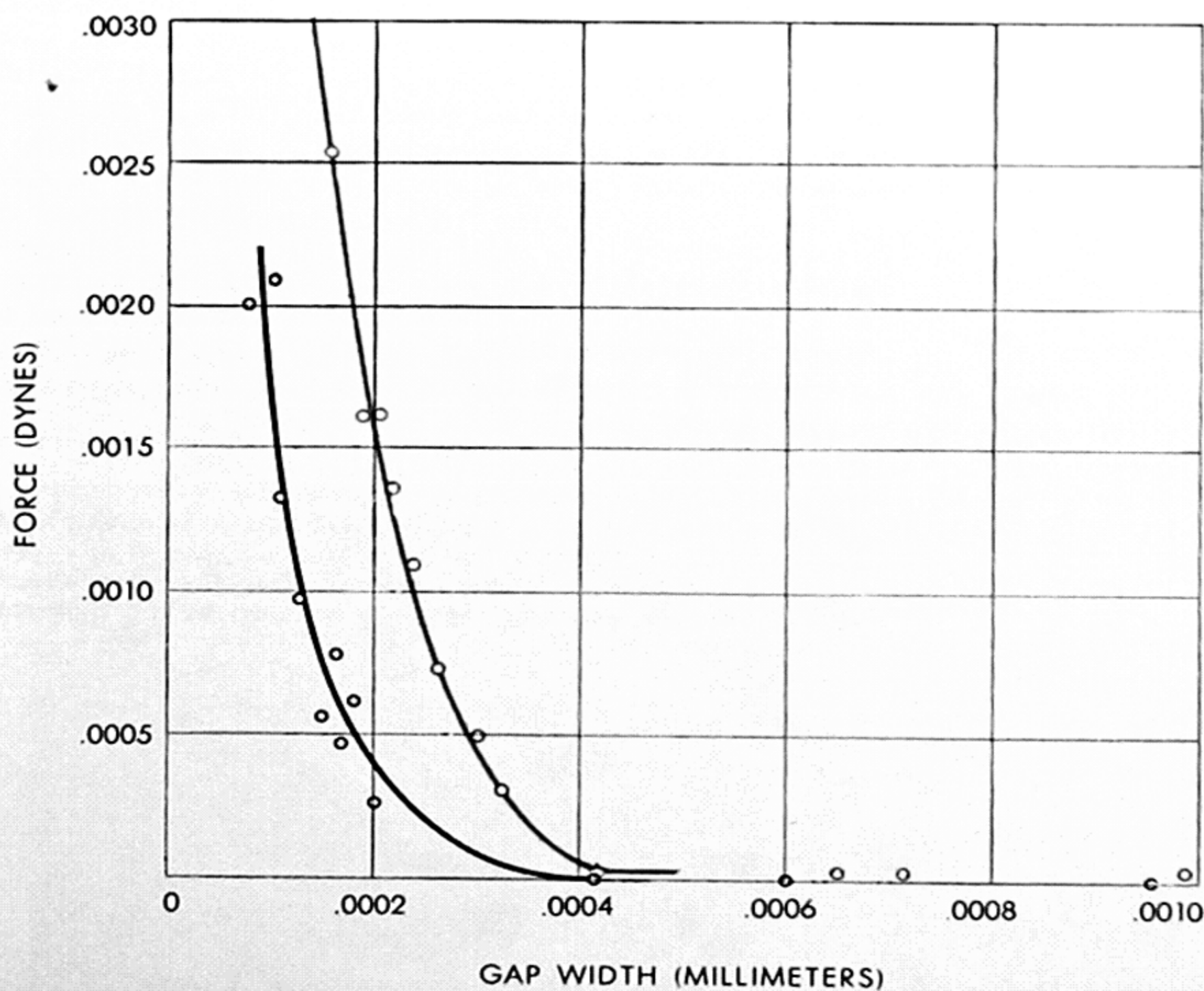
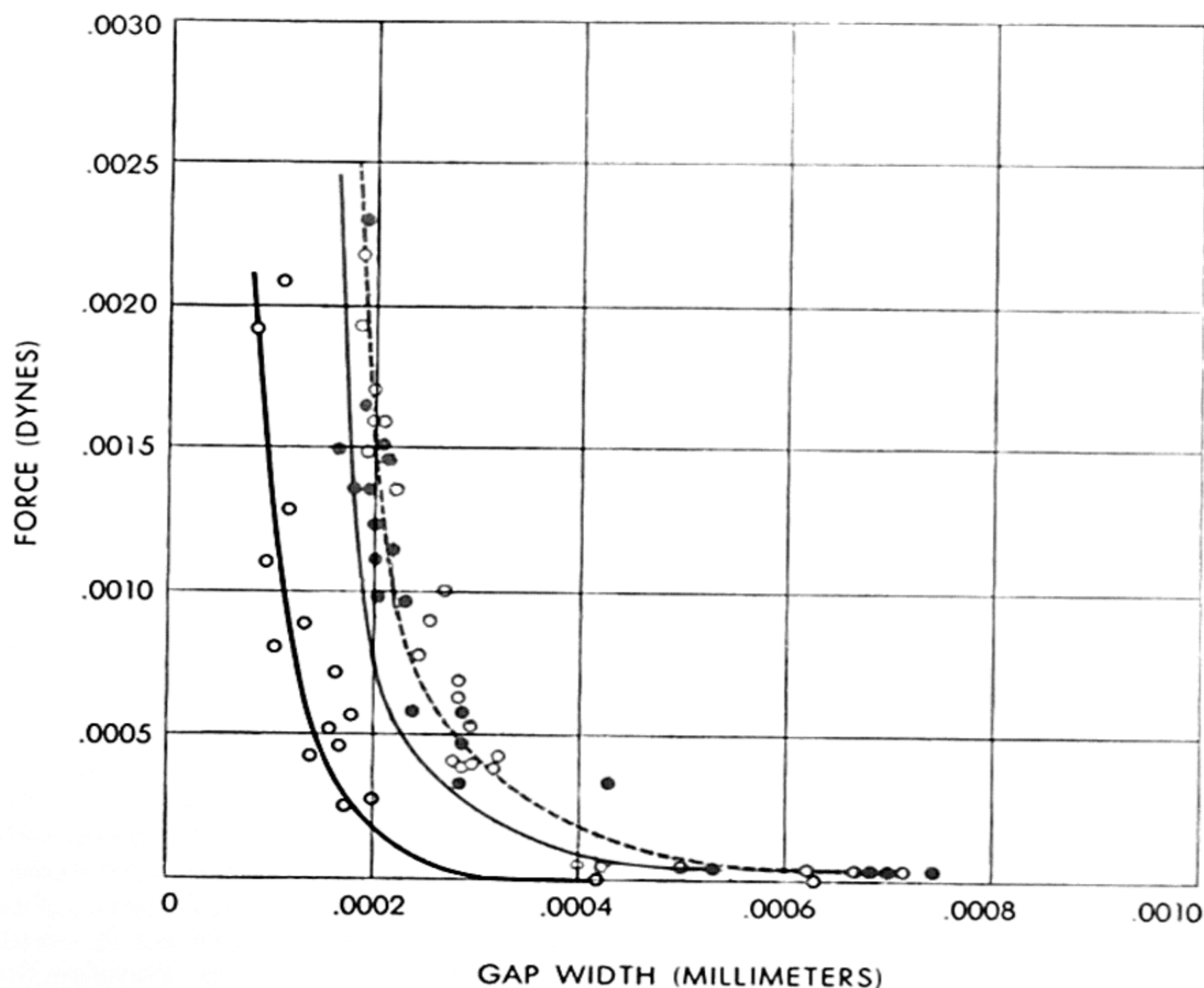
The theory is valid only for gap widths substantially greater than the molecular diameter.) But there is an even greater surprise. Whereas it is impossible, as we have seen, to pass mathematically from a microscopic, two-molecule theory to the macroscopic situation, the reverse transition is straightforward. The attraction between a pair of isolated molecules comes out of the Lifshitz theory as a special limiting case. For small distances it reduces to the London seventh-power law, and for larger distances, to the Casimir-Polder eighth-power law. And these equations are reached more easily by the roundabout method of Lifshitz than by the earlier direct calculations!

Furthermore, the calculations settle definitely the question of whether the force is additive. It is not. The attraction between two bodies is not the simple sum of the attraction between all pairs of their molecules. In fact, even on the macroscopic scale the force does not add [*bottom illustration on opposite page*].

In the case of macroscopic samples of matter the force depends on many factors, including the shapes of the bodies. When two flat, parallel plates are separated by a gap much smaller than the wavelengths of their main absorption lines, the complex formula reduces to a comparatively simple expression: The force varies inversely as the cube of the gap width. When the separation is much greater than the absorption wavelength, the force varies inversely as the fourth power of the distance. (These formulas correspond respectively to the inverse seventh- and eighth-power laws for pairs of individual molecules.)

As soon as Lifshitz announced his results, we applied them to a calculation of the force between the bodies we had used in our experiments. The curves at right represent the theoretical predictions. As can be seen, our experimental points cluster quite closely around the curves. This agreement between theory and experiment is particularly convincing because the theory contains no undetermined constants; the comparison is absolute and applies successfully to several different substances.

Although our experiments verify the Lifshitz formulas only for the special case of "large" gaps, they tend to support the whole theory. And so the electromagnetic nature of molecular forces and their relation to spectra, so long ago postulated by Lebedev, have now been demonstrated. He would be glad to know that the attraction between molecules does not "involve other forces whose source is still unknown."



EXPERIMENT AND THEORY are compared in these graphs. The curves show the variation of the molecular force of attraction with distance for the bodies used in the author's experiments, according to the Lifshitz theory. The points represent experimental values. The upper graph refers to experiments on a flat plate and a spherical lens 10 centimeters in radius. Data for quartz bodies are in black; for thallium halide, in solid color; for a quartz and chromium body, in broken colored line and open dots. Lower graph shows data for a 10-centimeter quartz sphere (black) and for a 26-centimeter sphere (color).

The Author

BORIS V. DERJAGUIN is the founder and director of the Laboratory of Surface Forces in the Institute of Physical Chemistry of the Soviet Academy of Sciences. He has held the title of Academician since 1946. From 1935, when he founded the Laboratory of Surface Forces, until 1941 he worked with Academician M. M. Kusanov investigating the forces involved in the coagulation of material suspended in solution. His studies have included the molecular theory of friction and the theory of adhesion of particles.

Bibliography

DIRECT MEASUREMENT OF MOLECULAR
ATTRACTION BETWEEN SOLIDS SEPA-

RATED BY A NARROW GAP. B. V. Derjaguin, I. I. Abrikossova and E. M. Lifshitz in *Quarterly Reviews*, Vol. X, No. 3, pages 295-329; 1956.

THE GENERAL THEORY OF MOLECULAR FORCES. F. London in *Transactions of the Faraday Society*, Vol. XXXIII, Part I, pages 8-26; January, 1937.

THE INFLUENCE OF RETARDATION OF THE LONDON-VAN DER WAALS FORCES. H. B. G. Casimir and D. Polder in *The Physical Review*, Vol. 73, No. 4, pages 360-372; February 15, 1948.

VAN DER WAALS FORCES. H. Margenau in *Reviews of Modern Physics*, Vol. 11, No. 1, pages 1-35; January, 1939.

ZUR THEORIE UND SYSTEMATIK MOLEKULARKRÄFTE. F. London in *Zeitschrift für Physik*, Band 63, Heft 3 and 4, pages 245-279; July 1930.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

THE STRUCTURE OF LIQUIDS

by J. D. Bernal

A new geometrical analysis shows that there is some order in the disorderly arrangements of the molecules of a liquid. The method may lead to a general theory of the liquid state.

“Structure” may seem an odd word to apply to liquids, implying as it does an enduring form. Is not lack of structure—fluidity—the very essence of a liquid? I shall argue here that it is not; that liquids do have, if only instantaneously, an internal molecular architecture in which the key to understanding their properties lies. In trying to get at their nature by way of fluidity it seems to me we are attacking secondary properties before learning to deal with the primary ones.

Whether or not I am right, there is no question that we do not yet understand liquids as well as we do solids or gases. The disordered state of the widely separated, darting molecules in a gas has been well known for about a century, since the classic researches of James Clerk Maxwell. The ordered state of molecules condensed in crystalline solids was taken for granted much earlier, and W. L. and W. H. Bragg finally demonstrated it by X-ray analysis in 1912. In both cases there are now theories that derive many of the properties of the materials from the relationships of their molecules.

In contrast our knowledge of liquids is largely empirical. Physicists and physical chemists have learned to predict fairly well the properties of any liquid from those of liquids already known. But the laws do not arise out of any fundamental theory, and they ignore molecular structure.

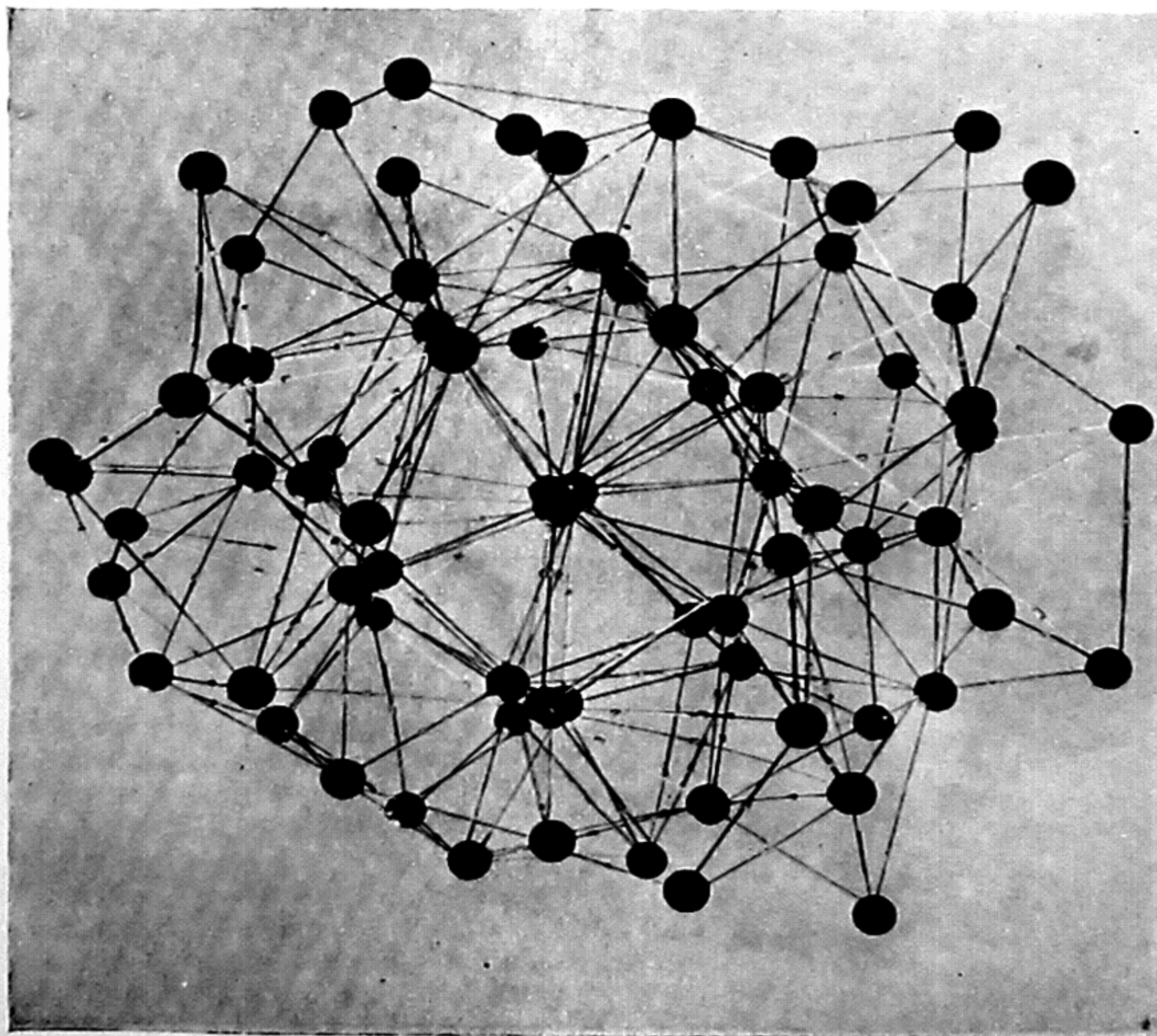
Attempts at a more fundamental approach have been frustrated by the hybrid nature of liquids. At once highly condensed and completely disordered, they present a very difficult mathematical problem. Simplified models, chosen because they lead to manageable calculations, are so implausible that even their inventors do not pretend that they repre-

sent the actual instantaneous structure of a liquid. If many of the calculated values based on them differ quite widely from experimental measurements, this is hardly surprising.

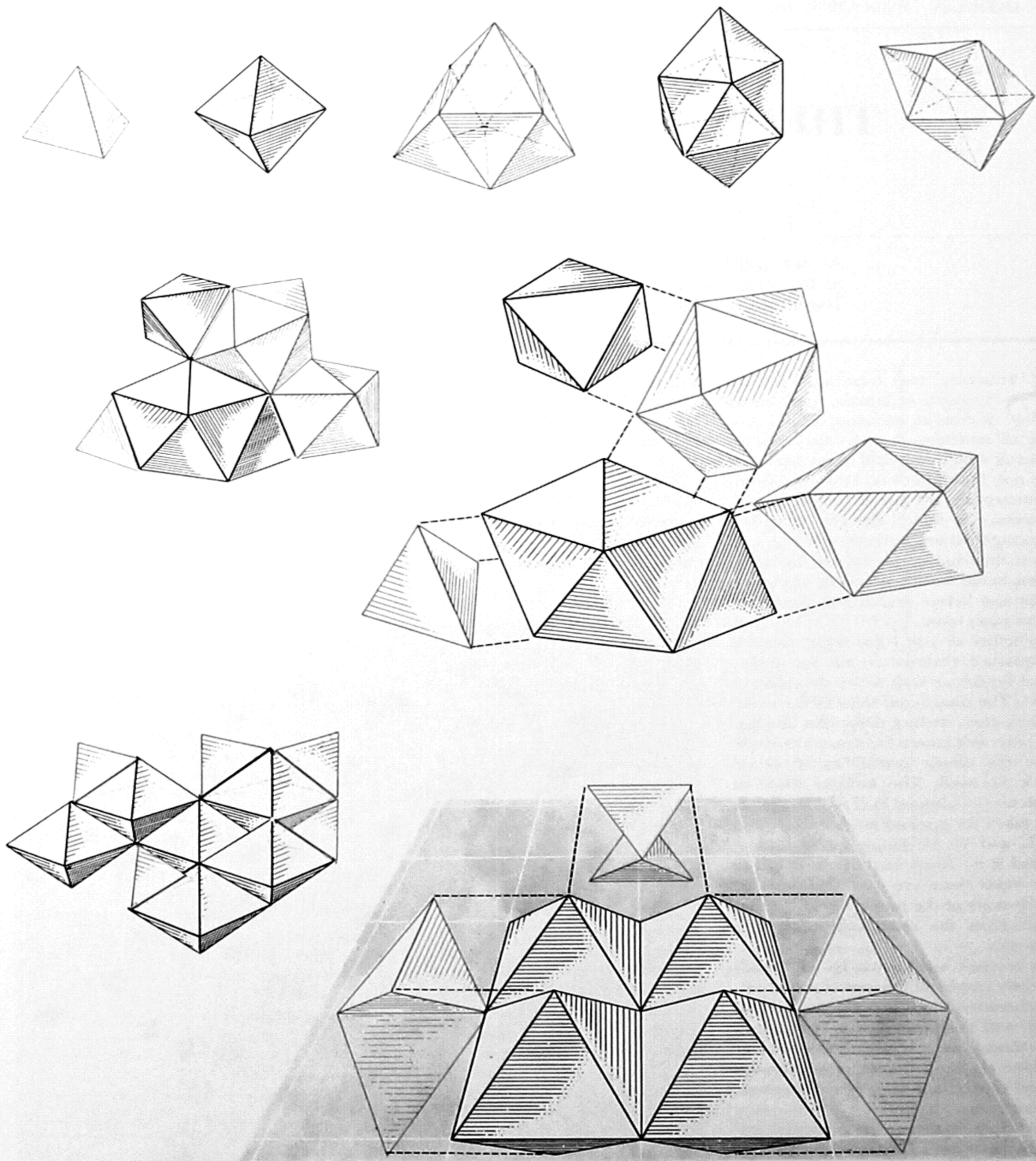
Nonetheless I believe that a molecular model, much closer to the actual arrangements, can be constructed and that, even if it is qualitative at first, it will be more likely to lead to correct quantitative

predictions. I have been working on the problem on and off for many years, more intensively in recent months. To put these efforts in proper perspective it will be helpful to consider a little more closely the relationship between liquids and the other two states of matter.

One primary property of a liquid is that it occupies a certain amount of space. At a given temperature or pres-



BALL-AND-WIRE MODEL of liquid structure is one of several built by the author. The position of each ball is limited only by the range of allowed distances to its neighbors. Such models can closely approximate the so-called distribution function of real liquids.



“POLYHEDRAL HOLE” MODEL of liquid structure, discussed in the text, makes possible a simple description of the irregular distributions of molecules in a liquid. In an ideal structure the molecules are all the same distance apart, and the lines between their centers form the edges of five polyhedra (*left to right at top*): regular tetrahedron, regular octahedron and three semiregu-

lar solids with 14, 16 and 12 faces respectively. Possible liquid structures correspond to various ways of fitting polyhedra together. Two arrangements are shown, nested at left and exploded at right. The upper pattern is typically irregular; if extended, it would not give a repeating arrangement. The lower pattern, with only tetrahedra and octahedra, is the only regular one that is possible.

sure it has a quite definite density, one that is affected much less than the density of a gas by a change in the conditions. Its bulk compressibility, as well as its density, is much like that of a solid.

In fact, when a liquid is not far from its freezing point, its structure cannot be very different from that of a solid. It occupies only about 10 per cent (in the case of molten metals 3 per cent) more room. Each molecule therefore must have about the same number of molecules surrounding it at about the same distance that it has when the material is in the solid state.

Not only that, but if the time-scale of observation is made small enough, the material actually exhibits solid properties. Ultrasonic vibrations of sufficiently high frequency set up shear waves in a liquid, as they do in a rigid solid. Experiments on the diffraction of neutron beams by liquids suggest that a molecule in a liquid has time to vibrate from 10 to 100 times before the structure changes. During that time the structure of a liquid is physically, though not geometrically, similar to that of a crystal. However, the irregularity of the liquid structure does allow a greater degree of tolerance. Crystals admit only a very limited degree of variation of composition. The use of crystallization to purify substances is evidence of the exclusiveness of the particular architecture that marks each crystalline phase. On the other hand, liquids mix much more readily, with each other and with both solids and gases. (They do not, to be sure, mix as readily as gases. Any two gases will mix in any proportion. Some liquids, such as the proverbial oil and water, will not.)

Hence it would seem that the atoms in a liquid do not occupy such closely specified positions as they do in solids; they have more elbow room, and they are not so particular about their partners. Indefiniteness or irregularity is an essential feature of the liquid state.

Another manifestation is the clear distinction between liquids—simple liquids, at any rate—and the corresponding crystalline solids. The melting point of a pure crystal is always a sharply defined temperature, but this sharpness of melting point appears only from the solid side. A liquid cooled through the freezing point, with appropriate precautions, does not solidify, nor do its properties change perceptibly when it passes through the temperature of freezing. In other words, the crystalline and the liquid phase are really two alternative ways of arranging molecules. They are at least as different



ANOTHER MODEL was made by placing spheres of plasticine in a rubber bladder and exhausting the air so that the balls were pressed together tightly enough to fill all the space.

from one another as the arrangements in different crystalline phases of the same substance.

There is, moreover, one other fundamental distinction between the structure of liquids and crystals. When a crystal is heated, the molecules vibrate and move farther apart, but do not change their neighbors. When a liquid is heated, on the other hand, both the identity and the number of neighbors change. A liquid therefore corresponds not to a single crystal phase but to a continuous series of such phases, each stable only for a single temperature.

When we compare liquids and gases, we find a much less marked distinction. True, at temperatures below the so-called critical temperature (above which a gas supposedly cannot be liquefied) a liquid has a definite boiling point, and a gas has the same condensation point at the same pressure. But the liquid can be markedly overheated and the gas can be undercooled. As the critical point is approached, the difference between liquid and gas fades out; at the critical point it appears to vanish. In my opinion this is only apparent; at temperatures well above the critical point it is still possible to demonstrate a sharp transition from a vapor-like to a gas-like phase, marked by a maximum in the specific heat (the amount of heat necessary to raise the temperature of one gram of the material one degree centigrade). According to my ideas, the essential difference between a liquid and a gas is that in a liquid the molecules are coherent:

every molecule is touching at least three others. In a gas the molecules are free or, at high pressures, are associated in small groups.

All this suggests what seems to me the simplest way of characterizing the states of matter in terms of molecular or atomic structure: Crystalline solids (all solids but glasses) have regular and coherent structure; liquids (including supercooled liquids and glasses), irregular and coherent structure; gases, irregular and incoherent structure.

Most theories of liquids have approached the problem from either of two extremes, considering liquids as disordered crystals or as condensed gases. Both viewpoints have led to useful conceptions, and even to calculations of properties, some of which agree reasonably well with experiment. But each is an unbalanced picture of a liquid, which is essentially disordered and essentially coherent. Furthermore, neither gives a detailed description of how the molecules are actually arranged, in short, of the structure.

There was an early attempt to get at liquid structure by means of X-ray diffraction, which had proved so effective in marking out the detailed arrangement of atoms in crystals. The X-ray diffraction pattern of a solid is a series of sharp rings, corresponding to reflections from crystal planes at various well-defined angles. The liquid pattern, however, consists of diffuse halos, usually no more than two or three.



COMPRESSED SPHERES from model at left are polyhedra of various irregular shapes. The most common number of faces is

13 and the most common number of sides to a face is five. In fact, almost every polyhedron has more five-sided faces than any other.

The latter pattern shows that the molecules in a liquid have no long-range order. That is, beyond two or three molecular diameters the arrangement in one place has no effect on that in another. At the same time the existence of the halos shows that there is some short-range order. This order is best described in terms of the so-called radial distribution function, which can be derived from the angular variation of X-ray scattering, and from which, in turn, several properties can be calculated. To understand what the distribution function is, imagine that we have picked a molecule (considered as a point) at random within the liquid and drawn a series of spheres around it with their volumes regularly increasing so that the volume interval between two neighboring spheres is always the same. Then the distribution function is simply the average of the number of other molecules between such neighboring spheres, at each distance from the central molecule. The distribution function is thus a measure of the average density as a function of intermolecular distance.

At very small distances the function is zero; molecules occupy a finite space and cannot be closer together than their diameters. The value jumps to a high maximum at the distance of the nearest neighbors of the molecule. It falls off, then peaks less sharply for the next to nearest neighbors and less sharply still for those at third remove. Very soon, however, it smears out to a uniform value [see illustration on next page].

To a first approximation the distribu-

tion function of a simple liquid resembles that of a dense but random arrangement of hard spheres such as marbles. Some years ago Joel H. Hildebrand of the University of California demonstrated this experimentally by suspending a number of gelatin balls in a liquid and shaking the container. Later B. J. Alder, also at the University of California, performed a similar experiment numerically, with a computer [see "Molecular Motions," by B. J. Alder and Thomas E. Wainwright; SCIENTIFIC AMERICAN, October, 1959].

What I have tried to do in the first place is to make a model of a liquid structure which will give a better approximation to the distribution function than the hard sphere model. To find it, I have tried to discern some order in disorder, some rules that govern the instantaneous arrangements of the molecules, irregular though the arrangements are. The eventual goal is to count the number of different possible arrangements. As we shall see, this would provide the basis for a general thermodynamic theory of liquids.

When I undertook the task, I began by assuming that a liquid consists essentially of a set of molecules similarly but never identically placed with respect to one another. I have restricted myself to the simplest case of spherical molecules, which corresponds to liquefied metals or to liquefied monatomic gases such as neon and argon. I have also assumed that the liquids are approximately

homogeneous. They may vary slightly in density from place to place, but not in general structure; there are no regions where the molecules are regularly arranged.

I spent considerable time trying to imitate this kind of irregularity with physical models, and found it fairly easy to do so. Among the constructions was one in which I built up an array of balls joined by stiff wires of various lengths, doing the job in my office, where I was interrupted every five minutes or so. This enabled me to achieve almost perfect randomness, because by the time I got back to work I had forgotten what I had been doing last.

The model turned out to have the right sort of disorder, and also approximately the right density. It occupied about 15 per cent more space than an array of "close-packed" spheres. I wondered, however, whether I might not have introduced some order unconsciously, so I checked the model against randomized arrangements of spheres, some of which are illustrated on these pages. What appeared in all these cases was the prevalence of five-fold arrangements among the balls surrounding any one of them. Now to a crystallographer this was somewhat shocking, for we had got into the habit of considering it as an axiom that molecules can be arranged with two-fold, three-fold, four-fold or six-fold symmetry, but that five-fold symmetry is not allowed. However, the rule actually applies only to crystals. It is true that one cannot form regular pat-

terns with five-fold symmetry that fill space solidly and extend indefinitely in three dimensions. It is like trying to pave a floor with five-sided tiles.

Only in the last few years have we appreciated that this is not a law of nature but a definition of crystals. In a non-crystalline, irregular structure five-fold arrangements are, for purely geometrical reasons, likely to be the rule. The reluctance of crystallographers to contemplate such arrangements accounts, I believe, for the fact that there have been so few attempts to understand the geometry of liquid structure.

In order finally to eliminate any human or physical elements that might have introduced order into the model, I turned to a purely mathematical method, using a computer at the University of London. My son, M. J. M. Bernal, devised a program for producing a dense but absolutely random distribution of points with the one condition of a minimum distance between them. Starting with a point as a center, the computer

picks at random a second point not less than the minimum distance away, then a third not less than the minimum distance from each of the first two and so on until no more will fit in a given volume. It turns out that the space around the center out to about three times the minimum distance can be filled with about 70 or 80 points.

The distances as measured by the computer could be used directly to set up a radial distribution-function. The result turned out to be similar to that of actual monatomic liquids, but not identical. In particular, the first peak was too broad, indicating that the distances between nearest neighbors varied more than in a real liquid.

The difference, it seemed clear, arose from the fact that the simple minimum-distance rule used in constructing the model ignores the attractive part of the force that exists between molecules of a real liquid. At extremely close range there is a strong repulsive force, which falls off very rapidly with distance (it

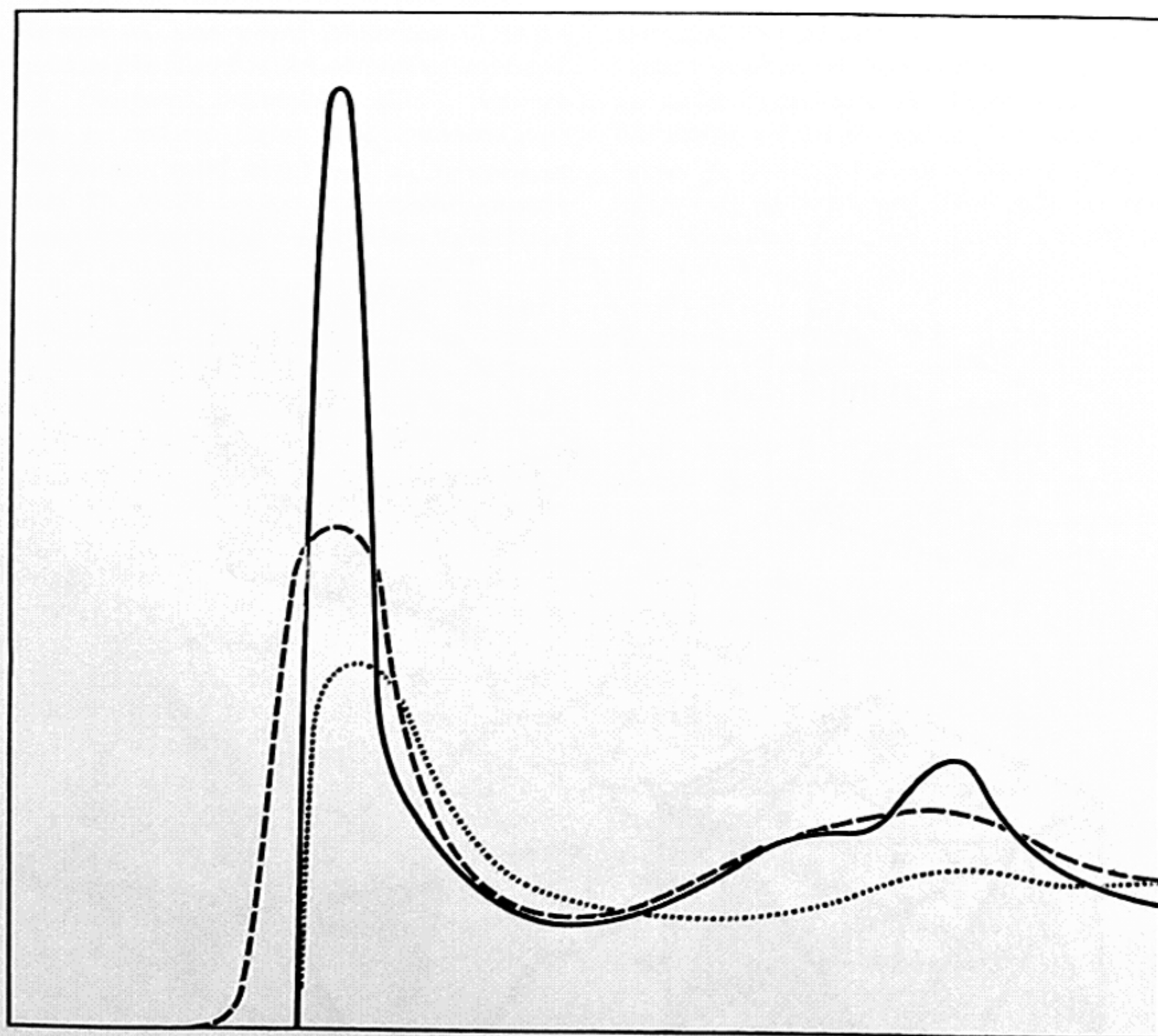
varies inversely as the 13th power). The curve is so steep that the molecules act almost as if they were hard spheres that do not repel until they touch. This part of the force is the one that the condition of minimum distance does represent. However, as the range increases, an attractive force appears. It, too, decreases rapidly (as the seventh power of distance), so that the force between non-neighboring molecules is effectively zero.

Immediate neighbors, on the other hand, are within the range of the intermolecular force. If nothing interfered, each pair would obviously lie at just the distance where the force disappears in changing from repulsion to attraction and where the relative potential energy of the molecules is zero. In an actual liquid the interactions of many neighboring molecules make the situation more complicated. However, there is a distance of separation, corresponding to minimum energy, at which the molecules tend to lie, and it is somewhat greater than the minimum distance of possible approach.

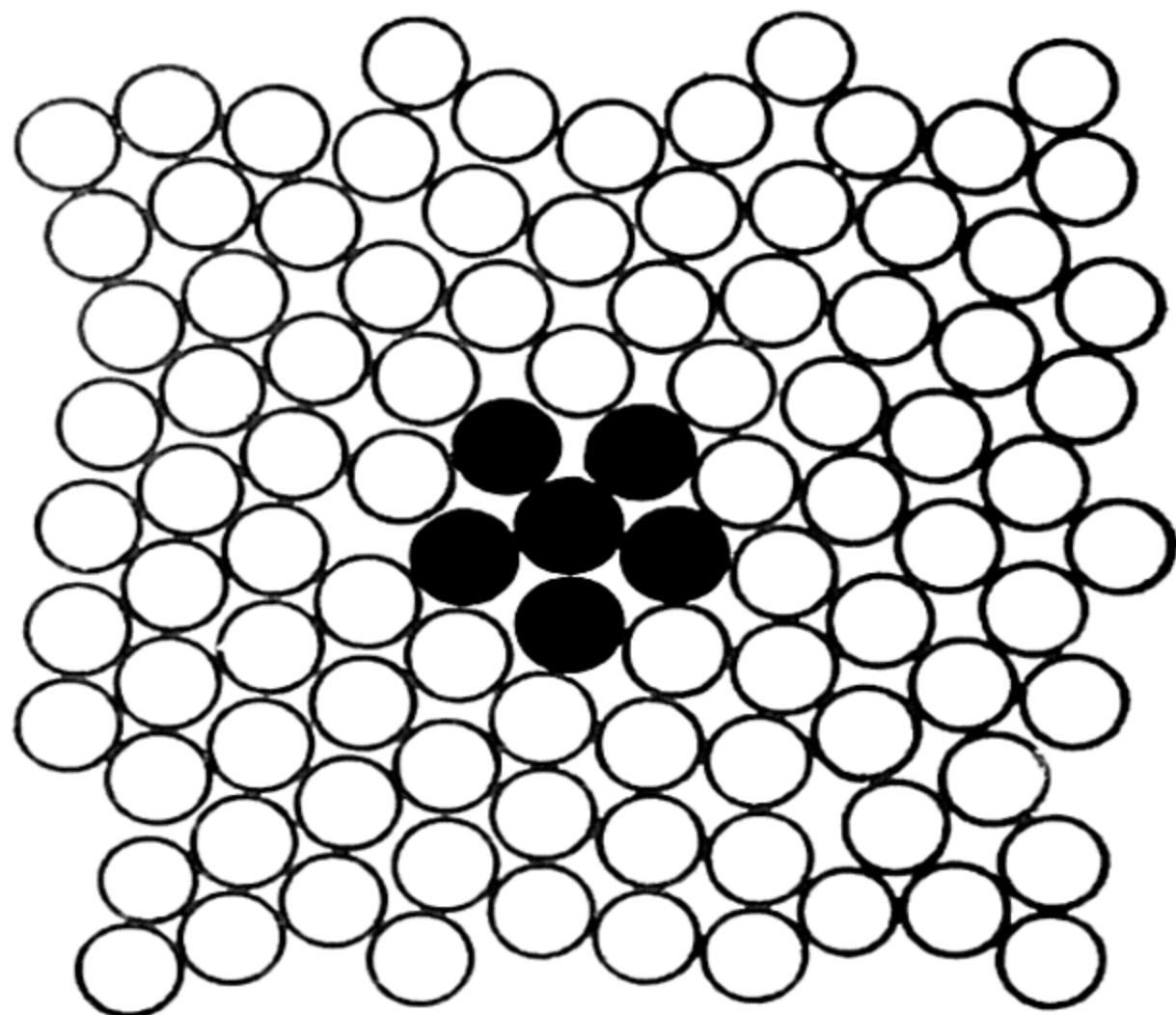
In order to take this factor into account, I built a ball-and-wire assembly corresponding exactly to the machine model, but with wires of adjustable length. Then I proceeded to squeeze the model until most of the distances approached the same value. By doing so I was able to come much closer to the actual distribution function. I was thus encouraged to think that the model, with all the neighboring balls now at approximately the same distance, represented fairly accurately the momentary situation of molecules in a liquid. The actual number of nearest neighbors of each molecule in the model varies from eight to 14, with an average of 11. In a real liquid this average coordination number, as it is called, determines the internal energy. When a liquid expands on heating, the average number of nearest neighbors of course decreases. Near the critical point, where the volume is three times the minimum volume, the number falls to between three and four.

It was only at this point that I saw that the equidistance model contained the key to the problem of finding the order of disorder. The answer was extremely simple, and I might have saved myself a lot of trouble by thinking hard first and computing and measuring afterward.

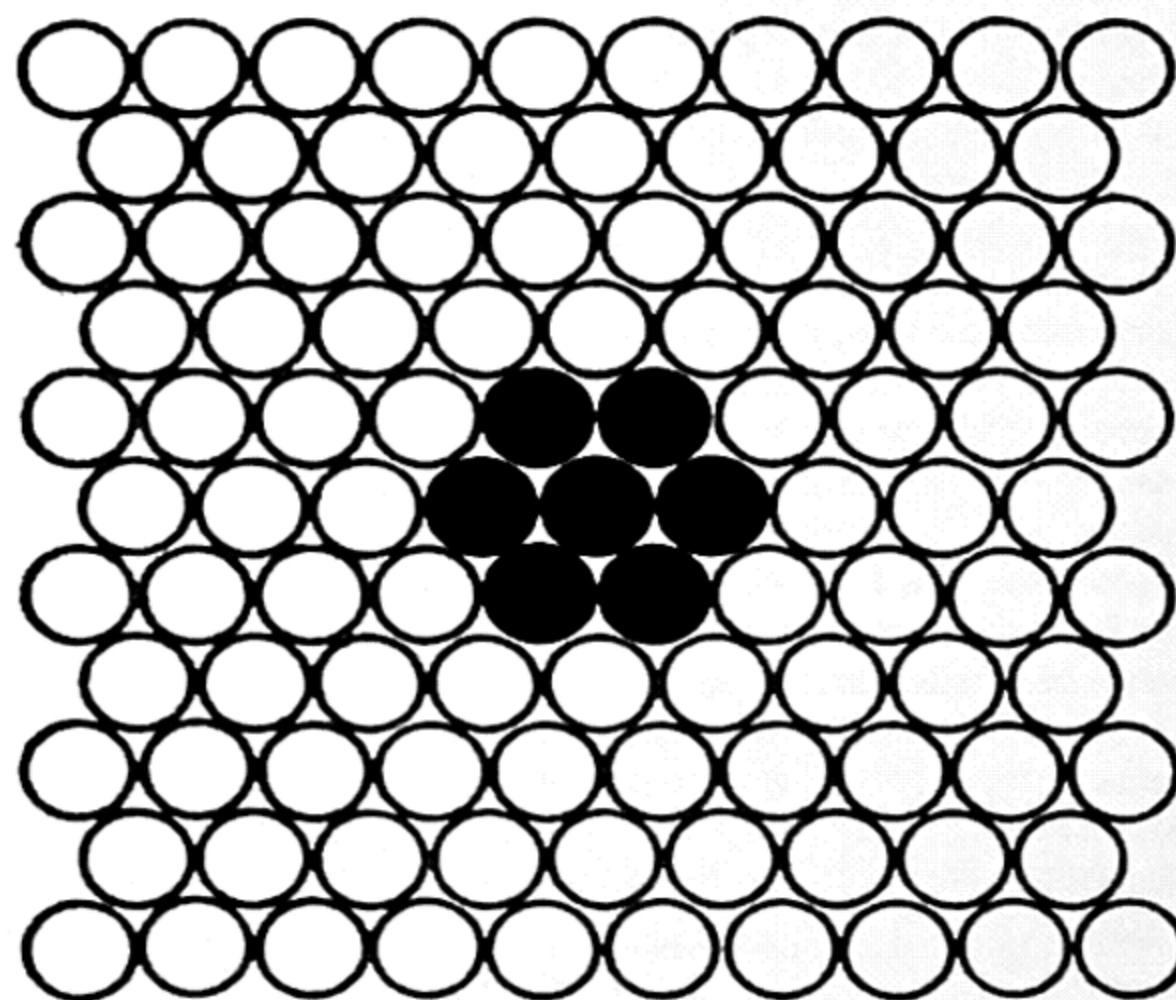
I conceived of an ideal model of a liquid, with each molecule surrounded by a limited number of others at equal distances. This corresponds to the ideal



RADIAL DISTRIBUTION FUNCTION of a liquid measures the probability of finding a molecule at any given distance from any arbitrarily chosen molecule, as explained in the text. Dashed curve represents the actual distribution function for molten lead, as determined by X-ray diffraction. Dotted curve is function derived from ball-and-wire model with balls placed at random but separated by a minimum distance; solid curve, function derived after adjusting model so as nearly to equalize distance between nearest neighbors.



FIVE-FOLD SYMMETRY, which turns up repeatedly in models of liquid structure, makes filling space with a regular array impossible. In two-dimensional representation at left the central



circle is surrounded by five others. The remaining circles cannot be made to form an orderly condensed pattern. Six-fold symmetry, as at right, leads to familiar regular hexagonal pattern.

model of a crystal from which the real crystal differs only in small atomic displacements.

Consider a point that is not at the location of a molecule. The molecules surrounding this point, or rather their centers, can be thought of as defining a hole in the shape of a polyhedron (a solid bounded by plane faces) with centers at the corners. In the ball-and-wire model the wires would represent the edges of these polyhedra.

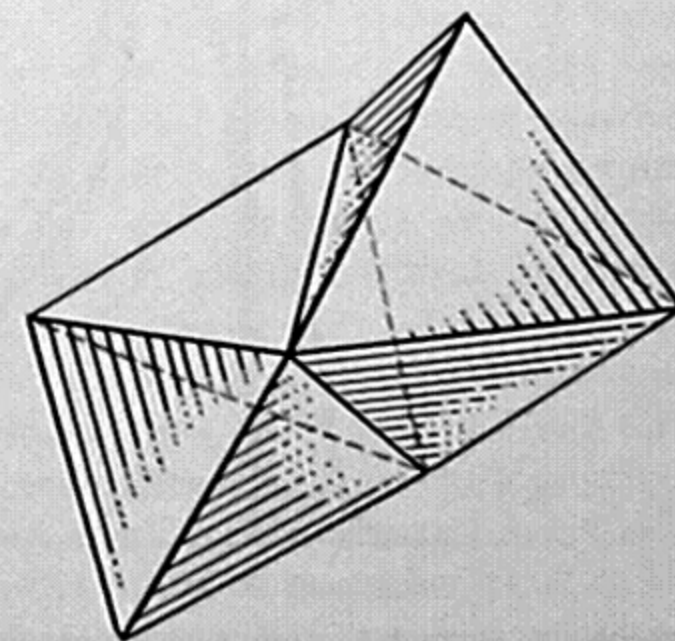
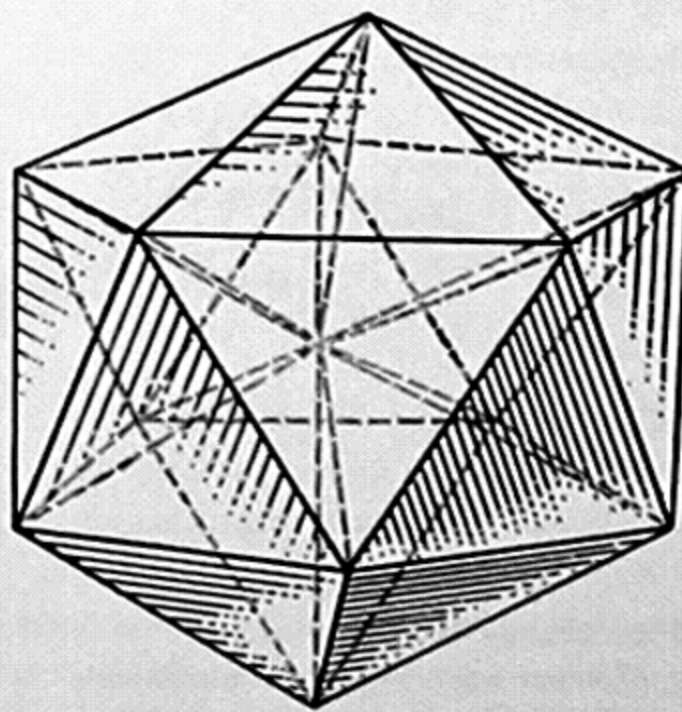
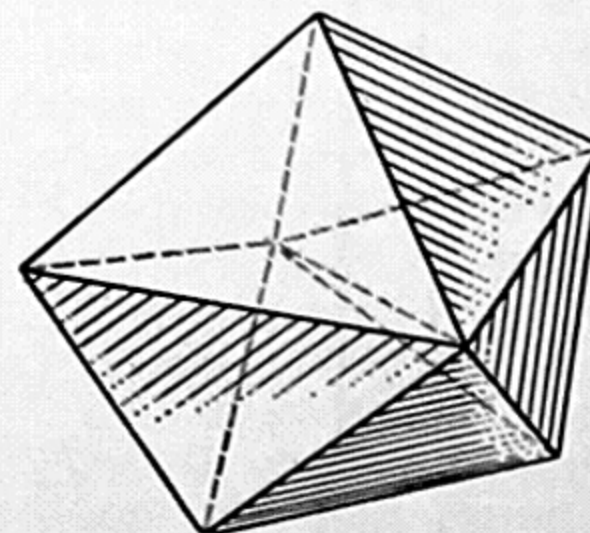
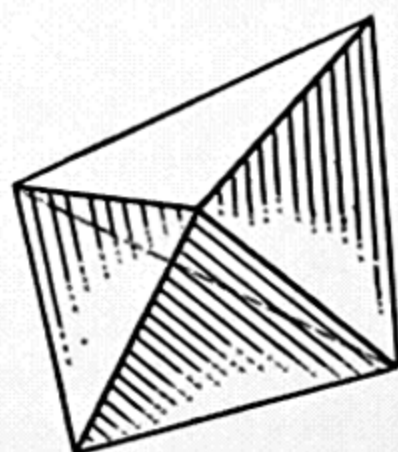
Now the significant thing about the ideal model is that all the edges of all the polyhedra are almost equal. If we limit ourselves to relatively dense packing, where there are no holes large enough to accommodate an extra ball, we need only consider the five smallest polyhedra with equal edges. These polyhedra, all with triangular faces, are the regular tetrahedron and octahedron, and three semiregular figures [see illustration on page 550].

Here is the key to the order we have been seeking. An ideal, dense-packed liquid is allowed only those structures that correspond to the ways in which some or all of the five polyhedra can be nested together to fill space completely. And the ideal model seems to be a reasonable approximation to a real liquid, at least to a liquid not too far from the freezing point.

When we examine how the holes go together, we find that the only combination that really fits is a mixture of tetrahedra and octahedra, in the proportion

of two to one. They can be put together to give two or three of the orderly arrays found in crystals. In any other selection the ideal polyhedra must be distorted slightly in order to fit together. A variation of about 10 per cent in the edge length is enough. Properly adjusted, the

holes fit together, but rarely in repeating arrangements. Most kinds of packing generate only indefinite irregular arrangements. The basic reason is the prevalence in the three larger holes of corners where five edges meet. As has been mentioned, five-fold symmetry can-



PSEUDONUCLEI are extra-dense aggregates of tetrahedra alone that are always found within close-packed arrangements of the five polyhedra of the polyhedral-hole model. The aggregates cannot grow indefinitely, however, because they eventually close on themselves.

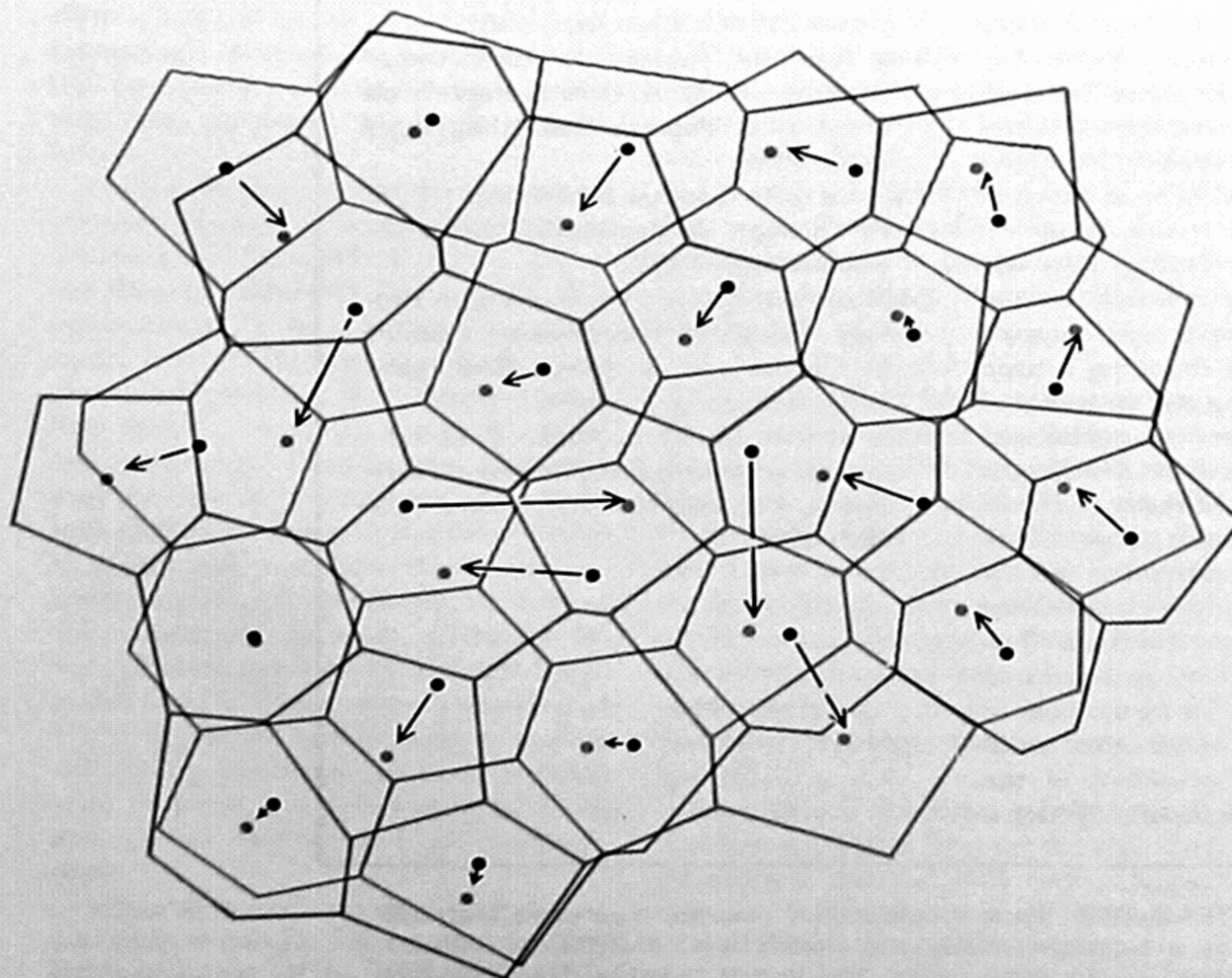
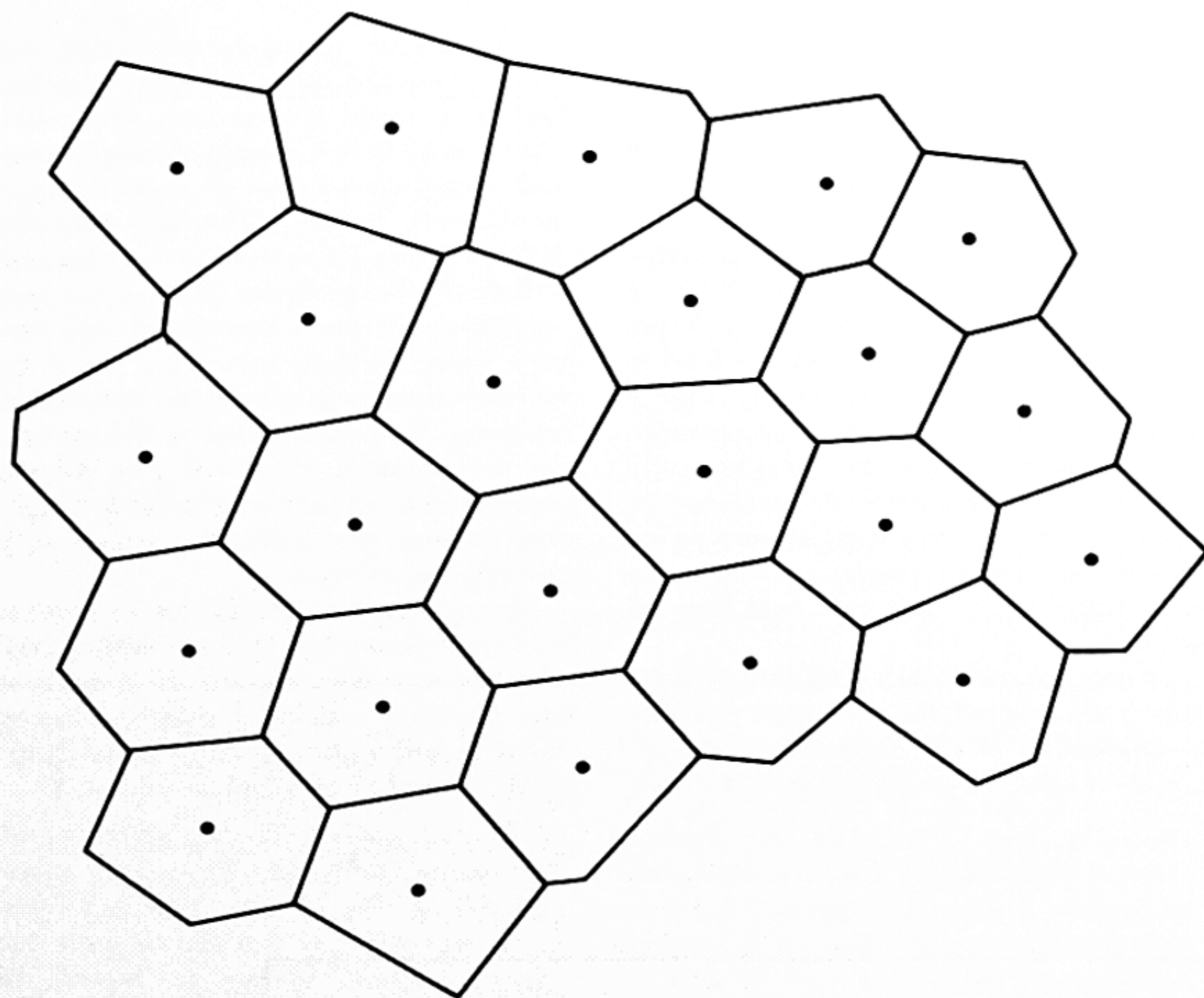
not lead to crystal-like structures.

Thus there is an enormous number of arrangements that correspond to the irregularity of a liquid, and only two or three that correspond to the regularity of a solid. In other words, just putting molecules together is very much more likely to lead to a liquid than to a solid. But the necessity for changing the length of the edges of polyhedral holes means that the molecules are not at their equilibrium distance, and consequently the energy is always high. This explains why every liquid will crystallize if the temperature is low enough. When some nucleation process starts the building of a regular arrangement, the arrangement always grows, because of its lower internal energy.

Closer inspection of the irregular arrangements corresponding to a liquid reveals that they have more tetrahedral holes than other kinds. Furthermore, the tetrahedra are not free and separated from one another by holes of other forms, but are mostly joined together in aggregates. These groups are very dense; even denser than the close-packed crystalline arrangements, since the latter must contain one larger, octahedral hole for every two tetrahedra. The super-dense aggregates cannot grow indefinitely, however. If you put three tetrahedra together and add another in one of the three possible ways, they join in a ring of five slightly distorted tetrahedra with one edge in common. Here again we meet five-fold symmetry—the form that prohibits any large-scale regular extension.

I have called these aggregates, which necessarily form in any dense irregular array, pseudonuclei. They are nuclei in the sense that they are hard and dense, in fact they are denser than true crystalline nuclei. But unlike them they are not viable; they can lead only to a very limited growth. Between the closed systems must be larger holes that can more than compensate for the extra density, and so the liquid arrangement is generally less dense than the solid.

The idea of pseudonuclei helps to explain the phenomena of supercooling and glass formation. As the temperature is lowered and energy is extracted from an irregular array, it may fail to crystallize, because the energy of snugly packed pseudonuclei is less than that of the same molecules in the regular array they adopt in the crystal. The reason why at low temperatures the crystal as a whole has lower energy, is that it does not have the relatively high-energy holes that must necessarily exist between pseu-



DIFFUSION IN LIQUIDS takes place as molecules continually change neighbors and shift among equally likely patterns. Structure at one instant (as at top and in gray at bottom) is shown as a two-dimensional diagram representing polyhedra around each molecule (not the polyhedral-hole model). Colored pattern shows structure at a later instant, after various pairs of polyhedra have lost or added common sides and after the molecules have moved.

donuclei.

The picture I have sketched so far is essentially static, but it provides a basis for explaining dynamic fluid properties as well. The essential feature of an irregular array is that there is at any temperature a number of arrangements differing by very small quantities of energy. To go from one to another is very easy, and this must occur spontaneously all the time. Thus a liquid has not one structure but a large number of equally likely structures, and is in constant flux between one and another. In each change-over some molecules change neighbors. The molecules move about in a random way; after a series of changes, original neighbors find themselves far apart; in short, the molecules diffuse.

Now if a stress is put on a liquid, those change-overs that tend to relieve the stress are favored, and those change-

overs that increase it are disfavored. The motion is no longer completely random, and the liquid as a whole flows. Thus we have at least a semiquantitative picture of diffusion and viscosity.

The only truly quantitative information to emerge from the model so far is the distribution function. My hope, however, is that it will eventually lead to a precise measure of disorder, or, in technical terms, to a calculation of entropy. Here is where existing theories go quite wrong. They fail, often by a factor of five or more, to give the right answer for entropy and for properties such as melting point and critical point that depend on it.

In essence, calculating the entropy involves finding all the different possible arrangements of the molecules and adding their relative probabilities. The low-

er the energy of an arrangement, the higher its probability. But the characteristic of a liquid is that very few molecules can be in low-energy arrangements, and then they can be in such arrangements only locally. The big difficulty with an irregular system is to recognize different arrangements and to be sure they have all been accounted for. Perhaps I can do it by analyzing the finite number of ways in which the polyhedral holes can be put together with one corner in common. This will give all the possible arrangements of nearest neighbors around any molecule, which may provide the solution.

My approach might be termed statistical solid geometry. It is too early to tell whether it will work out. If it does, it may not only provide a rigorous theory of the liquid state, but may prove helpful in a number of other problems.

The Author

J. D. BERNAL has since 1937 held the chair of physics at Birkbeck College of the University of London. Born in Ireland and educated in England, he studied physics at the University of Cambridge, began his research work under Sir William Bragg, and later lectured at Cambridge, where he was assistant director of research in crystallography. He has done most of his research on the structures of both simple and complex substances, and is now directing a team of workers investigating the structure of proteins, liquids, viruses, magnetic materials and corrosion products. During the war he worked for the Ministry of Home Security on protection against bomb damage, and was later adviser to the Air Ministry and scientific adviser to the Chief of Combined Operations. He is a Fellow of the Royal Society, which awarded him a Royal Medal in 1945, and a member of the science academies of six other countries. He is also vice president of the World Federation of Scientific Workers.

Bibliography

- KINETIC THEORY OF LIQUIDS. J. Frenkel. Dover Publications, Inc., 1955.
- THE KINETIC THEORY OF MONATOMIC LIQUIDS. G. H. A. Cole in *Reports on Progress in Physics*, Vol. XIX, pages 2-36; 1956.
- LIQUIDS AND LIQUID MIXTURES. J. S. Rowlinson. Butterworths Scientific Publications, 1959.
- MOLECULAR THEORY OF GASES AND LIQUIDS. Joseph O. Hirschfelder, Charles F. Curtiss and R. Byron Bird. John Wiley & Sons, Inc., 1954.
- THE STRUCTURE OF LIQUIDS. J. D. Bernal in *Proceedings of The Royal Institution*, Vol. 37, No. 168, pages 355-393; 1959.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

THE FLOW OF MATTER

by Marcus Reiner

Everything flows, including solids such as concrete and steel. The exact way in which a material flows is of practical concern to the technologist and a difficult problem for the physicist.

The mountains flowed from before the Lord," sang the prophetess Deborah (*Judges 5:5*). The English scholars who prepared the King James version of the Bible apparently could visualize only liquids as flowing, and they translated the Hebrew word for "flowed" as "melted." But the change was unnecessary. On God's time-scale even solid rock can flow. In the span allotted to us poor mortals the process is almost, but not quite, unobservable. By exceedingly accurate measurements we too can perceive rocks "creep" when they are subjected to large enough forces.

Indeed, we realize that the Greek philosopher Heraclitus was right when he declared that "everything flows." Today the flow of matter is the subject of a separate and lively scientific discipline. Apart from its interest as a challenging theoretical problem, the phenomenon of flow touches virtually every aspect of our technology. Construction engineers must be able to predict how soil will be squeezed away beneath the foundations of a building, and how steel and concrete beams will gradually stretch or warp under steady loads. Chemists design plastics that can be extruded into useful shapes. Bakers want a dough that works smoothly under their kneading and rolling machines. Airplane builders need to know how lubricating oil flows through an engine, and air across a wing.

Although a real understanding of flow has been achieved only quite recently, the beginnings of the story go back to the earliest days of physics. Isaac Newton was one of the first to consider the problem of fluid motion. An unconfined liquid (or gas) will flow whenever a force, no matter how small, is applied to it. However, the motion is resisted by

viscosity, "the lack of slipperiness of the parts of the liquid," as Newton put it. The equations of viscous flow were worked out by the French physicist Louis Navier and the English mathematician George Stokes early in the last century. Essentially they postulated that the rate of flow is directly proportional to the force and inversely proportional to the viscosity. For example, consider the three fluids air, water and castor oil. Water is about 50 times as viscous as air. If it takes one minute to pump a given quantity of water through a certain pipe, an equal pressure will push through the same amount of air in a little over a second. Castor oil, 1,000 times more viscous than water, would take almost a day to complete the same movement. In each case doubling the pressure would double the rate of flow.

As to solids, their most obvious property, and the first to be studied quantitatively, is not that large forces make them flow but that small forces do not. Under a small force a solid body is distorted, but when the force is removed, the body regains its original shape. Robert Hooke, a contemporary of Newton, discovered the fundamental law of this elastic behavior: that the amount of distortion is proportional to the force applied.

Of course if the force is big enough, the distortion goes beyond the so-called yield point, and the body can no longer recover. Such a permanent change of shape is known as plastic deformation, which is really another way of saying flow. In 1868 the French physicist Henri Edouard Tresca demonstrated that very high pressures could force metallic lead to flow through tubes. From these experiments Tresca's countryman, Barré de Saint-Venant, was able to develop

a theory of the plastic flow of metals.

Here, until 40 years ago, matters rested. Physicists and engineers thought in terms of two types of material: fluids, which flow no matter how small a force is applied to them; and plastic solids, which flow only if the force exceeds a critical amount known as the yield stress. The classification was apparently clear-cut. There seemed to be no difficulty in distinguishing between the flow of water, say, and of mild steel, or even of very viscous castor oil and very soft lead.

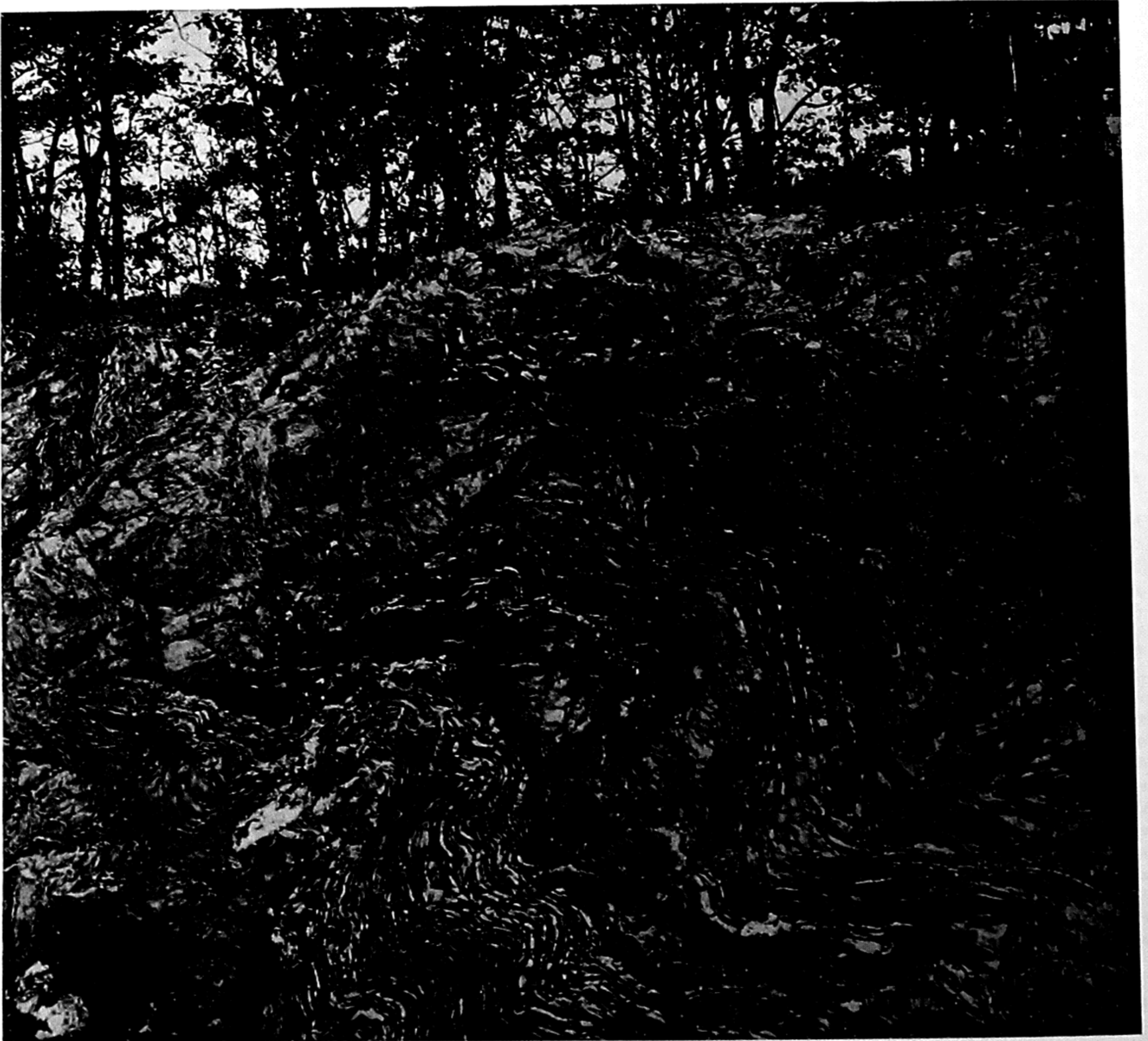
Then, in 1919, a seemingly routine problem in industrial chemistry thoroughly muddled this classical picture. E. C. Bingham of Lafayette College in Easton, Pa., was concerned with the properties of ordinary house paint. A good paint should spread easily under the brush, and should then flow enough to obliterate the brush marks. Consisting of pigments suspended in oil, paints were naturally considered viscous liquids. The requirements of "brushing" and "leveling" qualities evidently called for a low viscosity. However, a satisfactory paint must fulfill another requirement: brushed onto a vertical wall, it must not run off under the force of its own weight before it dries. This seemed to call for a high viscosity. Bingham came to realize that this apparent contradiction arose from a misconception of the problem. In order for the paint to stay put on the wall it was not enough that it flow slowly; it must not flow *at all*. In other words, it must be a solid! To be sure, it is a solid with very low yield-stress, so that the gentle pressure of the brush is enough to make it flow. Bingham published his findings under the title "Paint, a Plastic Material and Not a Viscous Liquid."

Surveying the classical ideas on flow, Bingham found them quite inadequate

to describe the behavior of substances like paint, and the new plastics, ceramics, lubricants, rubber products and so on that the chemical industry was beginning to produce. He and his colleagues then set out to develop a modern theory. At the same time, unaware

of these developments in the U. S., I was working on the same problem in faraway Palestine. Over the next few years we became acquainted with one another's studies, and a number of international meetings were held. A peculiar feature of the early meetings was

that, although they were attended mostly by chemists, the subject matter would have been more interesting to physicists. Finally, in 1948, it was decided to set up a new society for the study of deformation and flow, which would attract workers from all the allied fields. Named



FLOW OF ROCK is dramatically demonstrated by this formation laid bare by a highway excavation near New York City. These rocks originated with sediments deposited at the bottom of the sea in the

Older Paleozoic Period of some 400 million years ago. Between 20 million and 50 million years later, when the sediments had become rock, they were plastically folded as the sea bottom sank downward.

the Society of Rheology, after the Greek word for flow, it has just held its 30th-anniversary meeting at Lehigh University in Bethlehem, Pa.

Although the laws of rheology are essentially mathematical, they can be visualized with the help of some simple mechanical models. Rheologists themselves have formed the habit of thinking in terms of their models. This attitude, now largely out of style in physical sciences, harks back to the last century, when British physicists especially tried to picture abstract concepts such as the electromagnetic field and the ether by means of ropes, pulleys, springs and other such devices. Often these contrivances were considered to go beyond mere analogy; their inventors thought they represented the actual structure of the physical world. So also with rheological models. If we think of the molecules of a material as having elasticity, for instance, then a spring may not only be an analogue for the behavior of the material, but a real element of its internal structure.

In any case the models are built up from just three elementary mechanisms [illustration at right]. The first, called a Hooke body, is a helical spring rigidly supported at one end. It represents, of course, the elastic behavior of bodies. When a force is applied at the free end, the spring extends. When the force is removed, it goes back. The amount of the extension depends on the size of the force.

The second element is called a Newton body, and represents pure viscous flow. It is a "dashpot"—a tube filled with oil, in which a loosely fitting stopper can move as a piston. Pulling or pushing the stopper causes it to move through the liquid at a definite rate, proportional to the force. As soon as the force is taken off, the motion stops, and the plunger stays just where it is.

Finally there is the Saint-Venant body, a model of classical plastic flow. It consists simply of a weight resting on a horizontal tabletop. In order to move the weight the friction between it and the table must be overcome. A force too small to do this has no effect. A sufficiently large force moves the weight across the table. Again, when the force is removed, the weight comes to rest in a new position.

Actually the Saint-Venant body is a defective model. It accurately represents the phenomenon of yield stress, but not the true motion when the yield point is exceeded. A steady force large enough

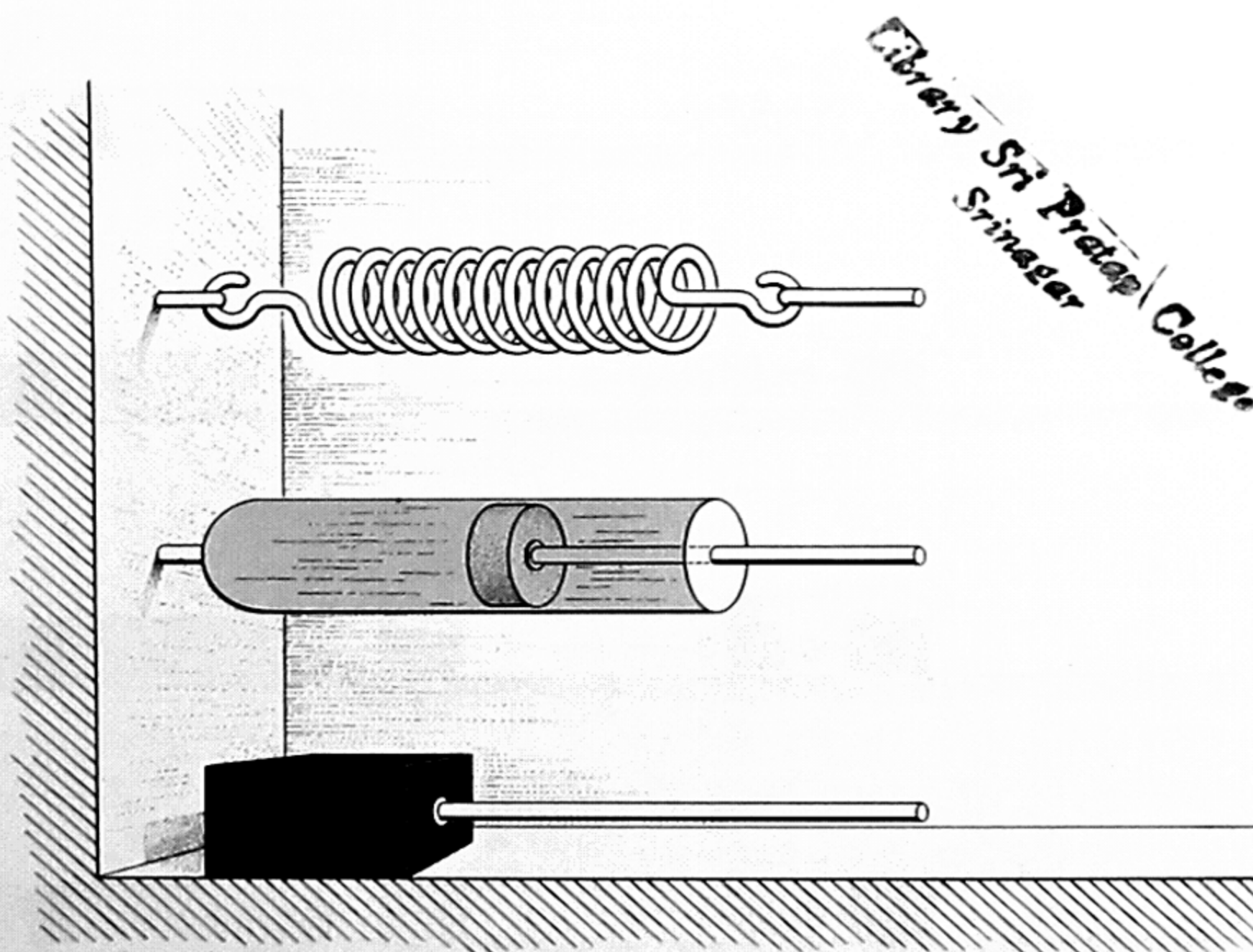
to set the weight in motion would in fact cause it to move faster and faster once it got started. But plastic flow, once it has been set up, resembles viscous flow in that a steady force produces a constant rate of movement. To remedy this shortcoming we can hook up a weight and a dashpot in parallel [see illustration at left on next page]. Then the weight determines the behavior below the yield point and the dashpot determines the motion of flow.

This is precisely what Bingham did in his analysis of paint, and the combination is known as a Bingham body. See how nicely it demonstrates the kinship as well as the difference between various materials. To represent paint we would envision a light weight with smooth surfaces, resting on a smooth table. The dashpot might contain a rather light oil. A very small force would be enough to start the body moving, and it would then flow quite easily. Lead, on the other hand, would be conceived as a heavy, rough weight on a rough surface, together with a dashpot containing very thick oil. Now the yield stress is high, and the flow comparatively sluggish.

As the reader will have anticipated, different combinations can be put together to suit a variety of problems. In brushing on paint, or squeezing tooth-

paste from a tube, or drawing wires from steel rods, we will probably care only about the properties of yield and flow, represented by the Bingham model. Suppose, however, we are concerned with the behavior of soil under a building foundation. Then its elastic deformation below the yield point would also be important. To represent it we hook a spring in a series with a weight [illustration at right on page 560]. A small force merely stretches the spring; a large one moves the weight as well. This arrangement is sometimes called a Prandtl body, after the German physicist Ludwig Prandtl.

Two more of the fundamental "bodies" of rheology were invented long before the discipline was born, by Lord Kelvin and James Clerk Maxwell. They did not describe the structures in terms of our elementary units, but it is possible to do so. Kelvin was thinking about the oscillations of metal wires. As a model he proposed a sponge made of perfectly elastic material, whose holes were filled with a viscous liquid. He pointed out that in a static experiment, such as stretching the sponge with a steady force, the model would be perfectly elastic: its elongation would be proportional to the force, and it would return to its original shape when the force was removed. On the other hand, in a dy-



BASIC MECHANISMS of rheological models are the Hooke body, a spring (top); the Newton body, a dashpot (middle); and the Saint-Venant body, a block that resists being pulled because of friction. Action of each body under the influence of a force is described in text.

Library Sri Pratap College

namic experiment—for example, rapidly stretching and compressing the sponge—the viscous flow of the contained liquid would absorb energy, and the body would behave like rubber or jelly. Kelvin's model turned out not to apply to oscillating wires, but it does describe

quite well a type of soil containing large grains of silt embedded in fine clay.

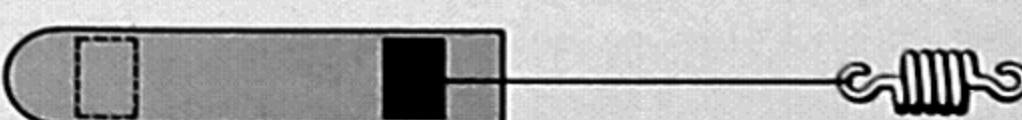
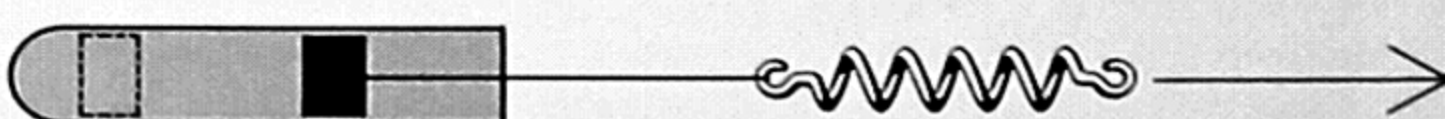
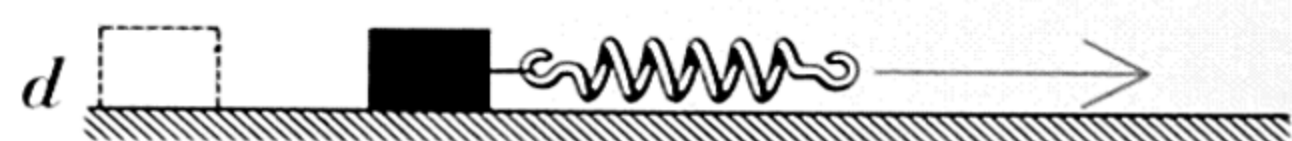
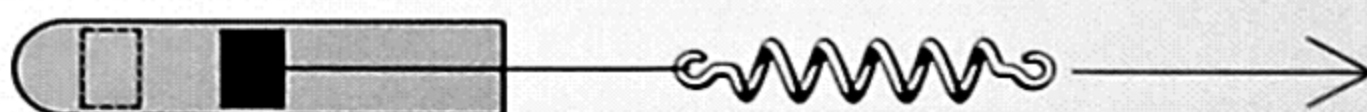
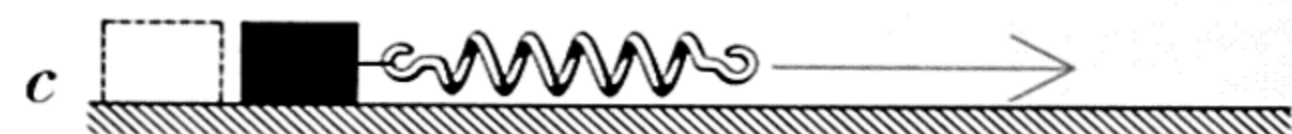
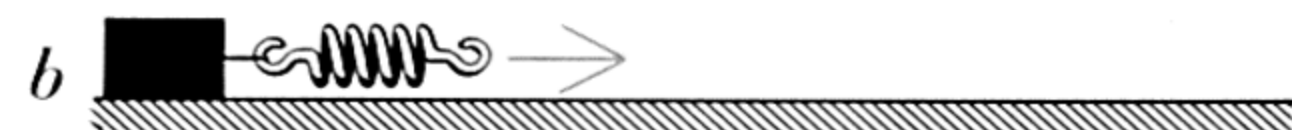
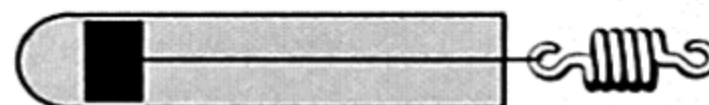
In rheological terms a Kelvin body is represented by a spring in parallel with a dashpot [see illustration 3 at right]. It is an elastic solid whose reactions are not instantaneous, but are retarded by

the viscous dashpot.

A Maxwell body also is made of a spring and a dashpot, but the two are hooked up in series [see illustration 4 at right]. It reflects the fact that fluids—even ideal gases, which Maxwell himself was studying—may have elasti-

1

2



MODELS FOR RHEOLOGICAL BEHAVIOR, made up of various combinations of the mechanisms depicted on the preceding page,

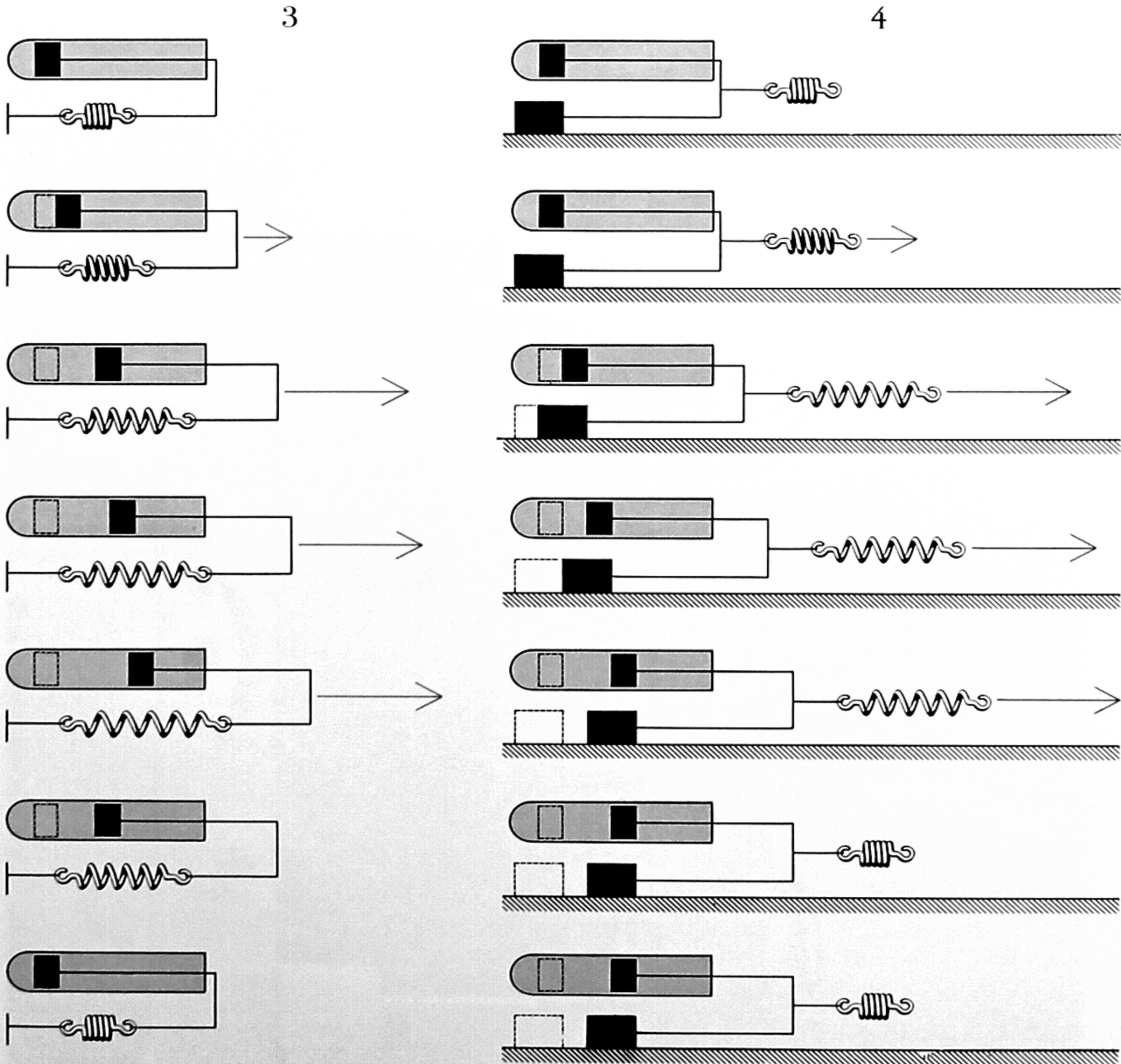
include the Bingham body (*column at left*), the Prandtl body (*column second from left*), the Kelvin body (*column third from left*)

ty as well as viscosity. If we pull on the model, the spring extends at once, and the piston begins traveling behind it. Thus we may think of the Maxwell body as a liquid that can be strained (*i.e.*, that can have its shape changed) elastically. If we keep the model stretched

at a certain length, the piston continues to move until the elastic strain disappears. We can say that the elastic stresses, or forces, in the liquid have "relaxed" in the time required by the piston to complete its motion.

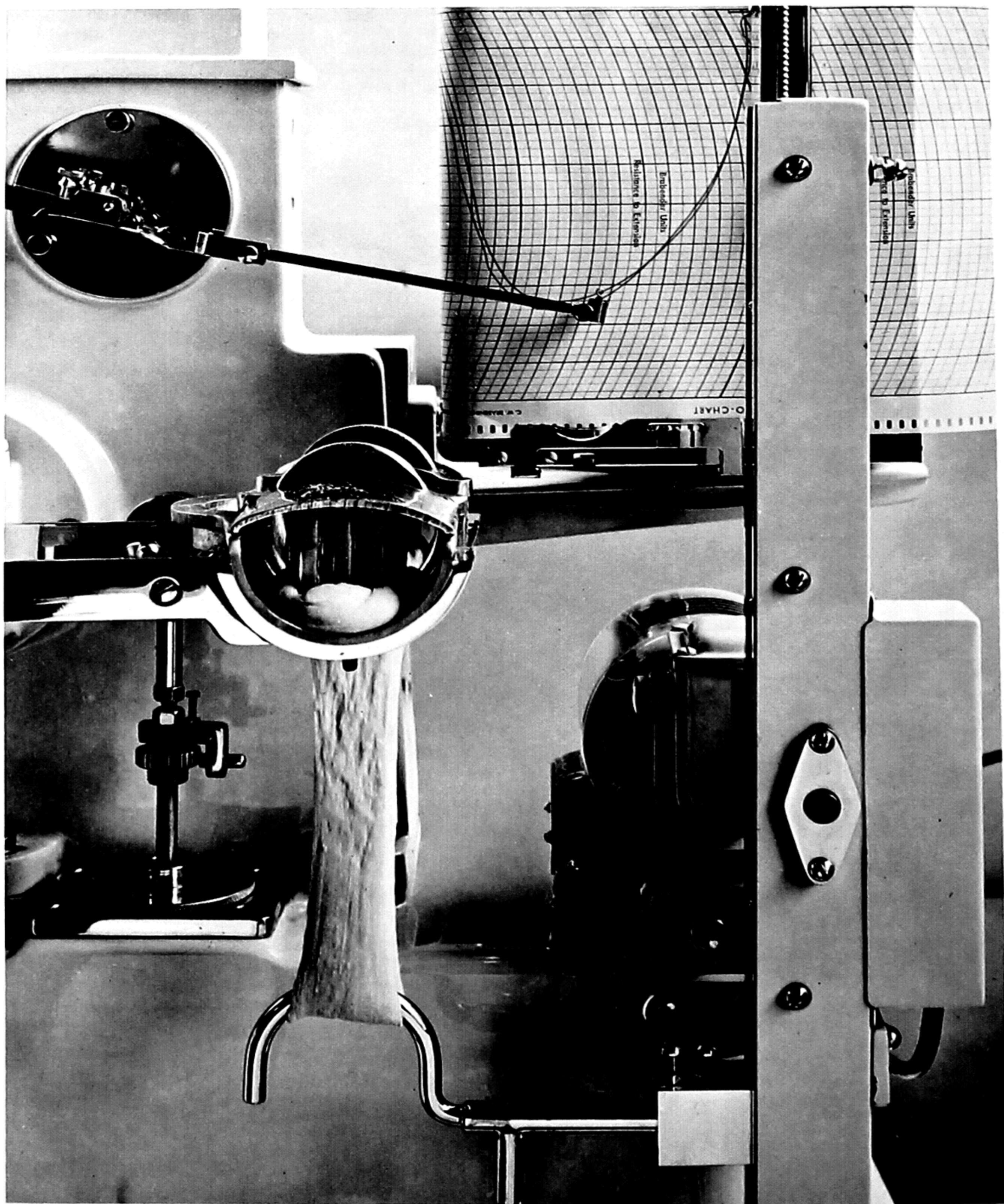
But there is another way of looking at

it. Imagine a rather weak spring attached to a dashpot containing extremely viscous oil. If the spring is pulled out and then quickly released, it will return almost exactly to its original position. The plunger will hardly have moved at all. Now the body looks like an elastic



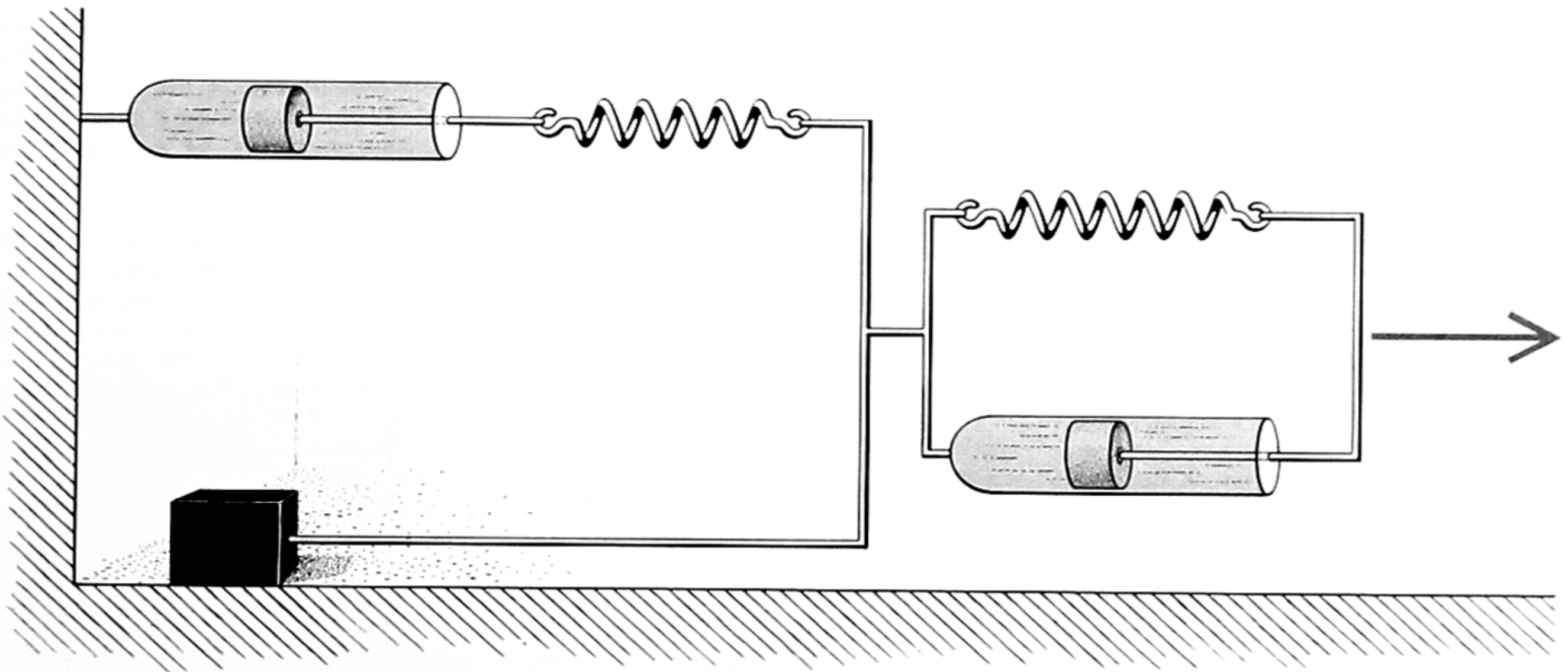
and the Maxwell body (column at right). Successive drawings in each column show the behavior of these bodies under the influence

of an applied force. Broken rectangles indicate original position of weight or piston; black rectangles, position at a given instant.



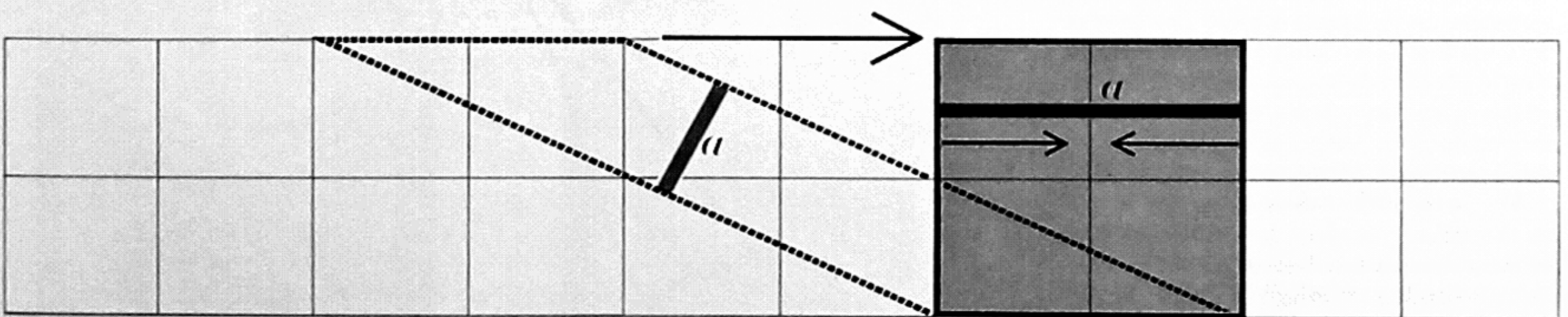
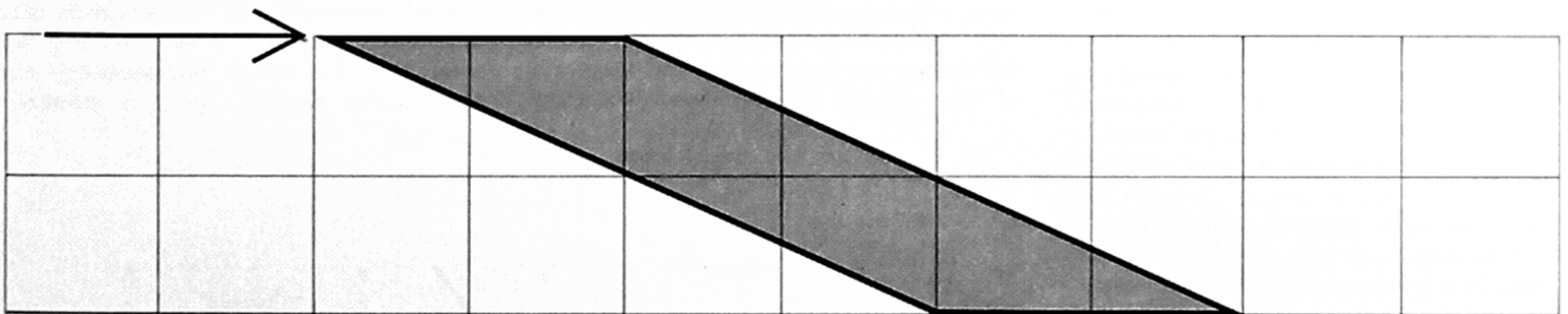
DOUGH IS MADE TO FLOW in "extensograph" of C. W. Brabender Instruments Inc. The dough is held in clamp in center and

stretched by hook at bottom. The curve on the graph at upper right indicates that the dough has good flow properties for baking bread.



COMPLEX FLOW OF DOUGH is represented by this intricate combination of rheological models. The model consists of a Max-

well body (*top left*) in parallel with a Saint-Venant body (*bottom left*), which are then connected in series with a Kelvin body (*right*).



SHEARING ACTION results in geometrically nonlinear flow. A block with slanted sides (*diagram at top*) is pushed by a shearing force to an upright position (*diagram at bottom*). The original dis-

tance between its sides, a , has now increased to the distance a . Thus the block is stretched sideways and is under tension, as is indicated by the two arrows that point inward from its sides.

solid. Depending on the viscosity we assign to the dashpot, it might be hours, weeks, even centuries before the fluid properties become apparent. From this point of view the everyday qualitative difference between a solid and a liquid becomes only a quantitative difference, depending on the relation between the

time of relaxation and the time of observation. If we could sit and watch Deborah's mountains for a few million years, we might also see them flow [see illustration on page 558].

By combining the basic models in various ways, extremely complicated types of rheological behavior can be

represented. The most elaborate case yet studied is the flow of dough. According to the British investigators R. K. Schofield and G. W. Scott-Blair, dough acts as if it were made of a Maxwell body in parallel with a Saint-Venant block, both joined in series to a Kelvin body [see top illustration above].

I leave it to the reader to picture for himself the strange movements that this model can perform.

Of course we do not suppose that dough is really the most complicated of all materials. On the contrary, every substance should be considered to have all possible rheological properties in greater or lesser degree. Their relative prominence will depend on the conditions under which the substance is examined. So far it is dough that has been observed under the most diverse circumstances.

The types of behavior we have considered so far, complex as they are, nevertheless share a certain underlying mathematical simplicity: they are all "linear." By this we mean that effects are directly proportional to causes. The two basic types of motion, viscous and elastic, are themselves assumed to be linear, as we have indicated. In elasticity the amount of stretch (or other change in shape) is proportional to force; in viscosity the rate of flow is proportional to force. The various combinations of these simple cases are also linear.

Unfortunately real materials are not always linear, a fact that first turned up in the study of liquids such as blood and solutions of rubber. Homogeneous liquids—water, alcohol, toluene, oil—and true solutions such as sugar in water have constant viscosity. Their rate of flow is directly proportional to the pressure. But when we force blood to flow through a tube, we find that doubling the pressure, for example, may triple or quadruple the rate of flow. In other words, its viscosity under large forces is less than under small forces. The plot of flow against force is no longer a straight line [*illustration at right*]. Liquids whose viscosity varies in this way are suspensions or dispersions—mixtures of different types of material.

We believe that their behavior is due to the effect on their internal structure of the deformation known as shear. Any liquid flowing through a tube moves faster in the center than near the walls. This difference in the rate of motion of adjacent layers is shearing action. In suspensions, shear presumably alters the internal arrangement of the constituents, making them flow more easily at higher rates of shear. Thus if the suspension consists of long, thin particles in a liquid, shearing might tend to align the particles in a common direction.

Whatever the exact mechanism, the phenomenon depends on the physical structure of the material, and hence is called physical nonlinearity. Liquids

that behave in this way are called non-Newtonian.

Although the basic model elements—spring, dashpot and block—are not ideally suited to represent a non-Newtonian liquid, it is possible to combine them in such a way that they will simulate its action. Thus it seemed until fairly recently that all possible types of rheological behavior could be analyzed in terms of the three basic elements. Then, during World War II, Karl Weissenberg in England discovered some totally new and strange effects (which I, at the same time, predicted from theoretical considerations).

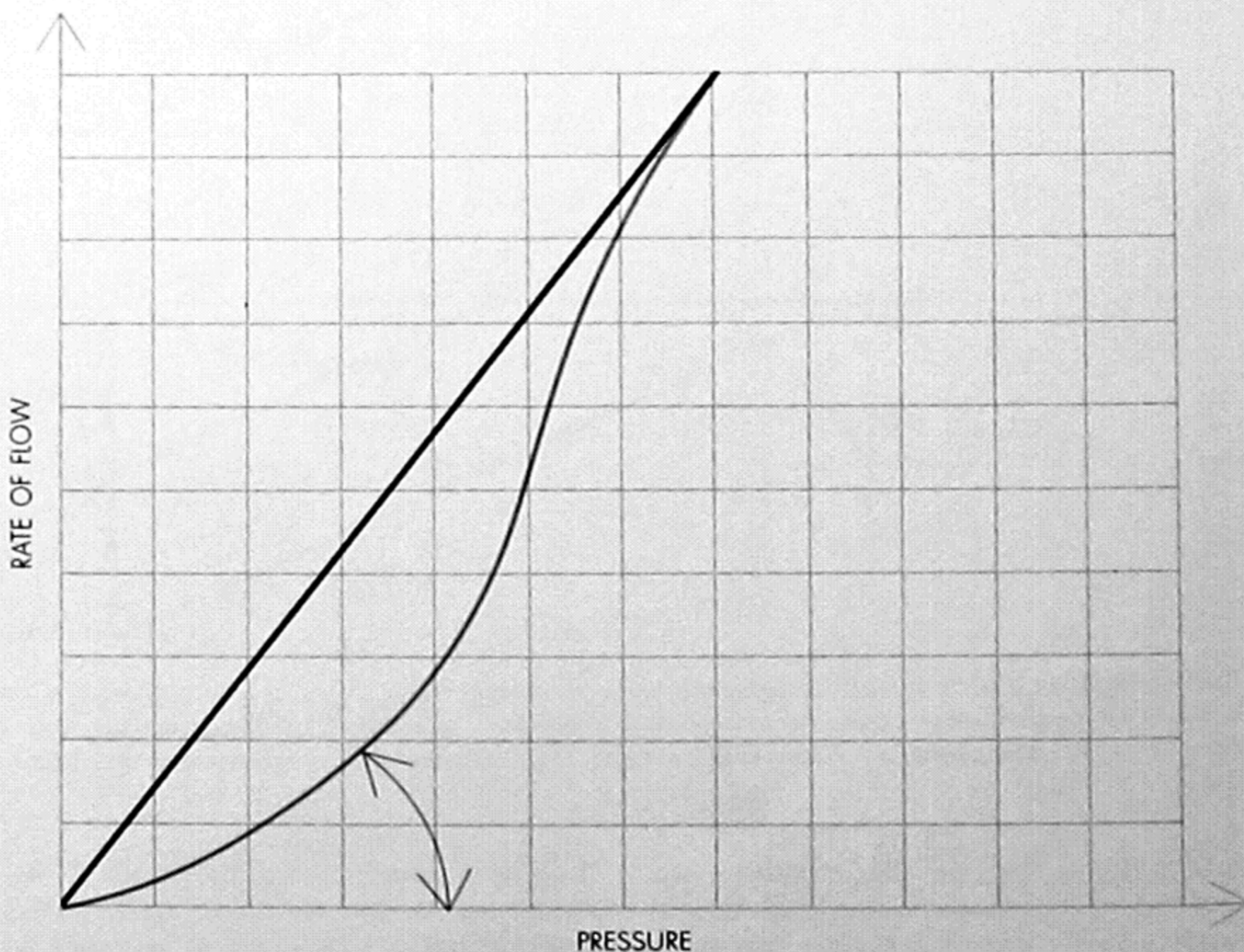
Experimenting with napalm and other viscous materials he saw these liquids do things that were simply impossible according to the standard theories. An investigator observing these motions for the first time, Weissenberg wrote, "would hardly believe his eyes."

We will mention just one example of the Weissenberg effects. Think of what happens when a rotating rod, like the shaft of a soda-fountain mixer, is lowered into a liquid. The liquid is set into rotation and, under centrifugal force, tends to move outward toward the walls of its container. The net result is to push

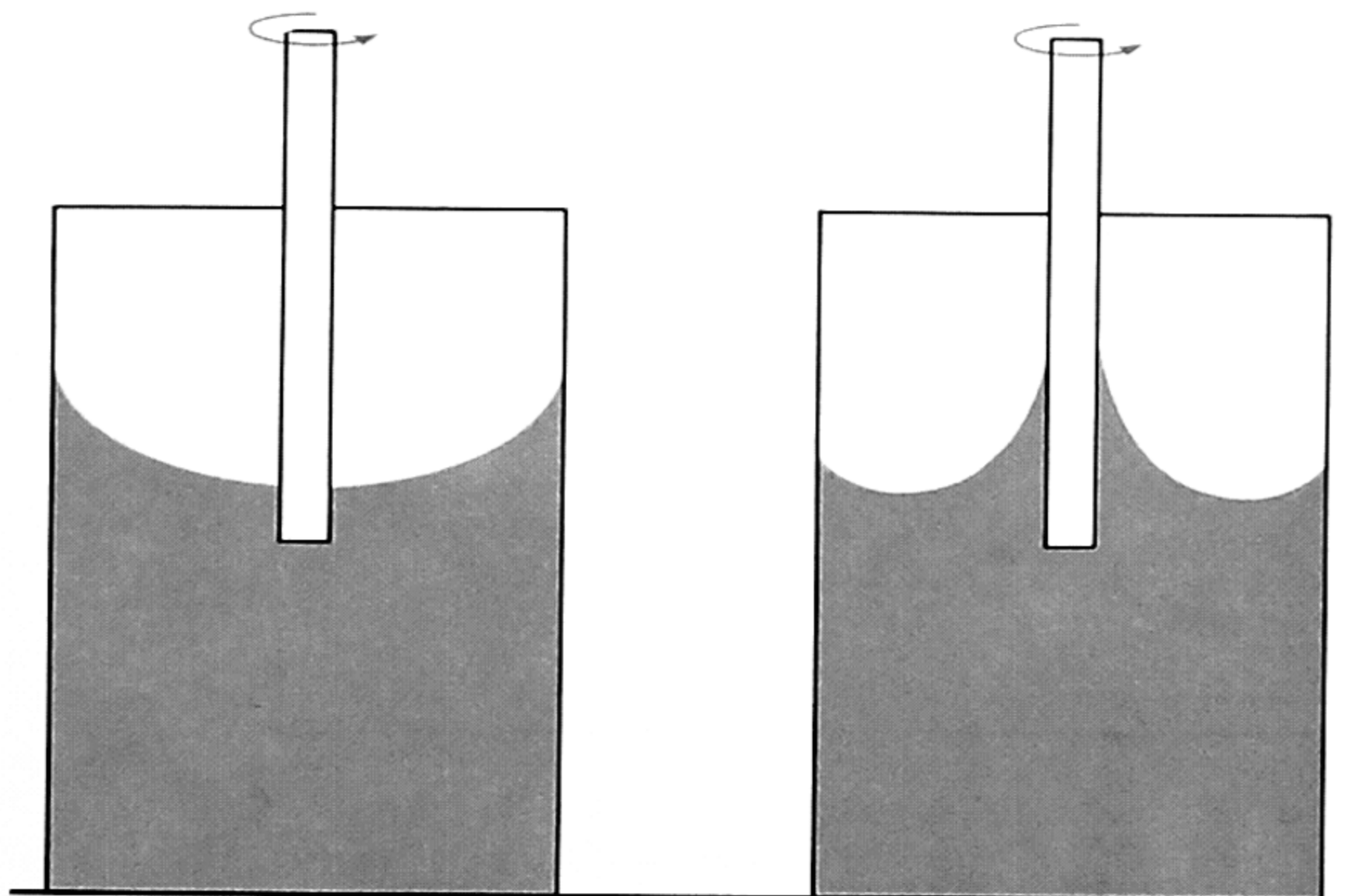
the liquid down in the center, around the rod, and up toward the walls [*see illustration at left at the top of page 565*]. When Weissenberg tried this [*illustration at right*], he saw his liquid climb up the rod!

The search for an explanation of such phenomena has demonstrated that some fundamental processes, supposedly well understood, are more subtle than they seem. For our purposes we need consider only the case of "simple" shear elasticity. Imagine an elastic block with slanting sides and a horizontal top and bottom [*see bottom illustration at left*]. The bottom is held fixed, and a horizontal force is applied to the top, pushing the block until its sides are upright. In describing the condition of the block, we now say it contains shear stresses that push back against the applied force and tend to bring the body back to its original shape.

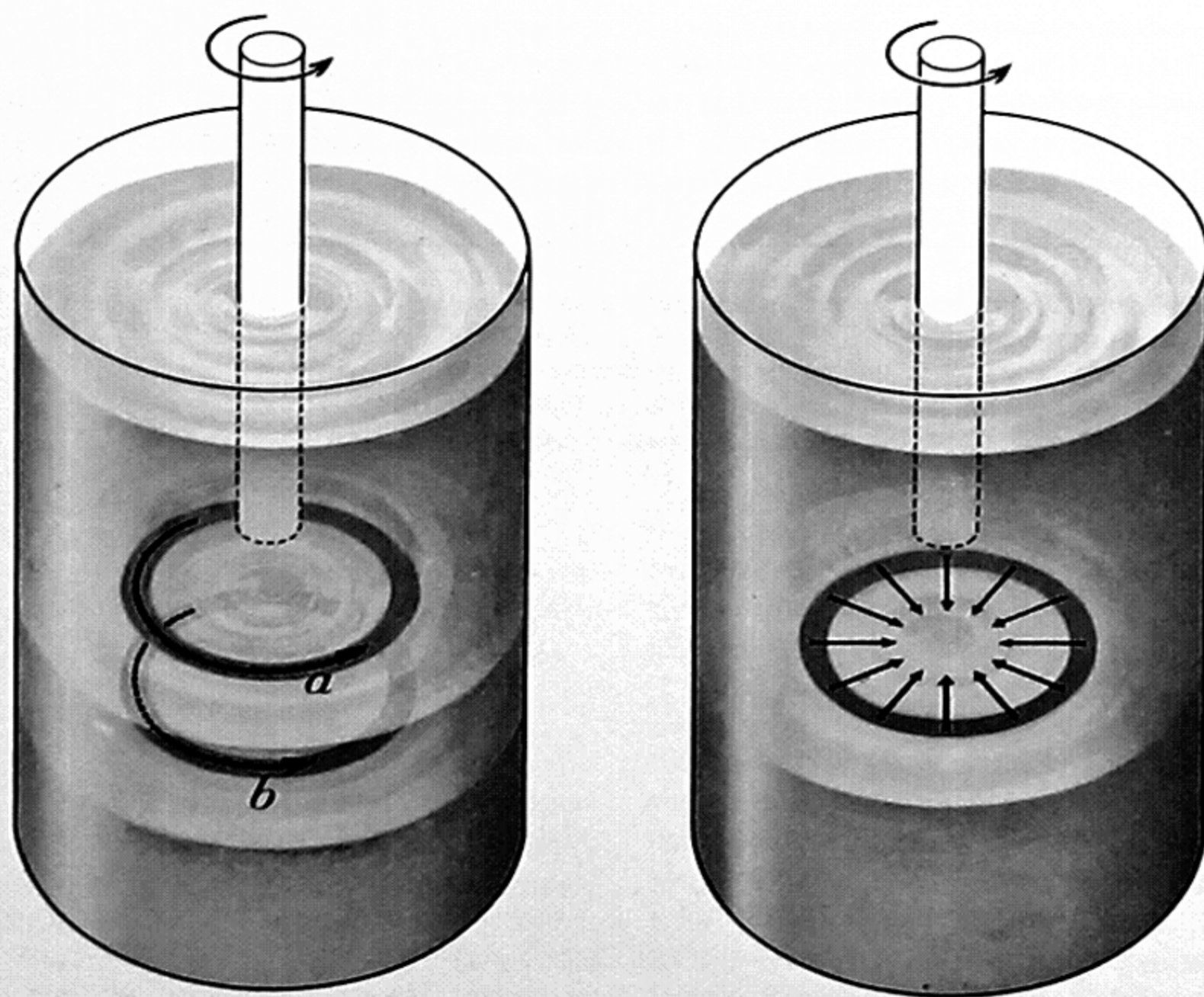
But there is more to it. If we compare the distance between the sides in the two positions, we realize that they are farther apart after the deformation than before. In other words, shearing has stretched the block horizontally, and it is under tension, as if a stretched



VISCOSITY PLOT of a Newtonian liquid is a straight line. The rate of flow is directly proportional to the pressure applied. A non-Newtonian fluid has a curved graph indicating that the viscosity of the fluid changes with different rates of flow (or different pressures).



WEISSENBERG EFFECT is illustrated by the drawing at right. At left a Newtonian liquid reacts to circular stirring by moving down at the center of the vessel and up near the walls. A liquid obeying the Weissenberg effects climbs up the stirring rod rather than down.



WEISSENBERG EFFECT is explained by shear elasticity. Rotation of liquid is faster in plane *a* than in plane *b* (left); therefore the liquid is sheared circularly. Because of elasticity the circular flow-line (right) is under tension along its length. Hence it acts like a stretched rubber band and tries to squeeze the liquid in toward the center of the vessel.

spring connected its sides. The tension is parallel to the original shearing force.

Now a liquid set into rotation by a rod is sheared, because the higher layers tend to move faster than the ones below [see illustrations on this page]. Here the shear is circular. If the liquid has elasticity, there will be tension in the same direction as the shear, as there was in the block. Therefore a circular flow-line acts like a stretched elastic band, tending to shrink down, forcing the liquid inward and up the rod.

This effect is not linear. It depends on the geometry of the deformation; hence it is called a geometrical nonlinearity. We now realize that, like other rheological properties, geometrical nonlinearity is shared by all materials to some extent. We can find it if we know how to look.

Recently I discovered a dramatic confirmation of this. If Heraclitus declared that everything is fluid, we can now say everything is solid—even air! In brief, I have found that air too has shear elasticity and can be made to act in the same way as the liquids we have just been discussing. The experimental setup consists of two parallel circular plates, one spinning and the other stationary. The stationary plate has a hole in the center to which is connected a pressure gauge. When the plates are about one millimeter apart, the air between them, set into rotation, tends to fly out because of centrifugal force. The gauge registers a drop in pressure. Now if we move the plates slowly closer, the pressure is gradually restored. At last, when the gap is only about a hundredth of a millimeter, the air is squeezed inward and through the hole, and the pressure jumps above normal.

The disks make a new kind of air pump. Moreover, if there is no hole in the stationary disk, the compressed air forms a very low-friction bearing. Both of these applications are now under development.

It is fair to say that not all authorities agree with my interpretation of this effect. Some have attempted to "save" the Navier-Stokes equations and to derive the result for a strictly Newtonian fluid with no shear elasticity. I believe, however, that my case is now pretty well proved. Thus rocks are liquid, air is solid and things are not what they seem.

The Author

MARCUS REINER has since 1949 been professor of applied mechanics and head of the Israel Institute of Technology. He was born in Austria and holds the degree of doctor of technical sciences from the Technische Hochschule of Vienna. Before World War I he designed bridges in Austria, and during and after the war he headed a special Austrian army unit for the reconstruction of steel bridges. In 1922 he joined the Department of Public Works of the British Government in Palestine, and devoted two decades to the design of reinforced-concrete structures. He visited the U. S. in 1928 and again as a research professor at Lafayette College and a lecturer at Princeton University in 1931. He made

a third visit to the U. S. this year. He has published many technical papers in structural engineering, hydrodynamics, theory of elasticity, flow of homogeneous liquids and air, and applied mathematics.

Bibliography

INDUSTRIAL RHEOLOGY AND RHEOLOGICAL STRUCTURES. Henry Green. John Wiley & Sons, Inc., 1949.

THEORY OF ELASTICITY AND PLASTICITY. Harold Malcolm Westergaard. Harvard University Press, 1952.

A TREATISE ON THE MATHEMATICAL THEORY OF ELASTICITY. A. E. H. Love. Dover Publications, Inc., 1944.

TWELVE LECTURES ON THEORETICAL RHEOLOGY. Marcus Reiner. North-Holland Publishing Company, 1949.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

THE NUCLEAR FORCE

by Robert E. Marshak

The deflection of high-speed particles by samples of matter is yielding new information about the strong and complex force that holds the atomic nucleus together.

What holds the nucleus together? The problem remains one of the most challenging in physics, as it has been ever since Lord Rutherford discovered that the atom has a nucleus. Discussing the question in *SCIENTIFIC AMERICAN* six years ago [September, 1953], Hans A. Bethe guessed that it had consumed "more man-hours than have been given to any other scientific question in the history of mankind." Since then the man-hours have continued to pile up, and considerable progress has been made.

It would be satisfying to be able to report an advance in fundamental understanding. The fact is that we are not noticeably closer to the eventual goal: a theory that will allow us to deduce nuclear forces mathematically from a few basic assumptions and experiments. But we do know much more about what the forces are like. While still unable to predict them, we are at least learning how to measure them.

This is no trivial accomplishment. An enormous effort has gone into designing and performing experiments to illuminate different aspects of the complicated interaction of the particles that make up nuclei. And once the experiments were done, theoretical physicists had to work equally hard to interpret the results.

Simpler Forces

Before plunging into the complexities of nuclear forces, let us recall, for purposes of comparison, some of the properties of more familiar, and much simpler, forces. Consider first the force of gravity. As everyone knows, it is an "inverse-square law" force: the gravitational attraction between two masses varies inversely as the square of the distance between them. (If the distance

is doubled, the force is divided by four, and so on.) Gravity is also a "central" force. It depends on the distance between the attracting masses but not on their relative direction, and always points along the line joining the two masses [*see illustration at left on next page*].

There is another way of looking at the interaction between two attracting (or repelling) bodies that the physicist often finds more convenient. Instead of talking about forces he speaks of potential energy. The potential energy stored up in a pair of attracting bodies is equal to the work that would be necessary to pull them infinitely far apart, so that they would no longer act on one another. Obviously the amount of work depends on the way in which the force of attraction decreases with distance. For gravity, where the force varies as $1/r^2$, the potential energy varies as $1/r$, r being the distance of separation [*see illustration at top of page 570*]. Generally there is a simple relation between the two concepts: Force is measured by the rate of change of potential energy. In graphical terms this means that where a plot of potential energy is steep, the corresponding force-curve has a large value. Where the potential-energy curve is comparatively flat, that is, has a low rate of change, the force is small.

Suppose now that we did not know the laws of gravitational force or potential energy. How could we discover them? One possibility would be to study the motions of objects under the influence of gravity. The laws of motion tell how bodies react to any force. Thus by observing their particular paths and speeds when acted on by gravity alone, the nature of the force can be deduced. When Isaac Newton calculated the elliptical paths of the planets around the sun, he was proving his inverse-square

law of gravity, as well as demonstrating the validity of his laws of motion.

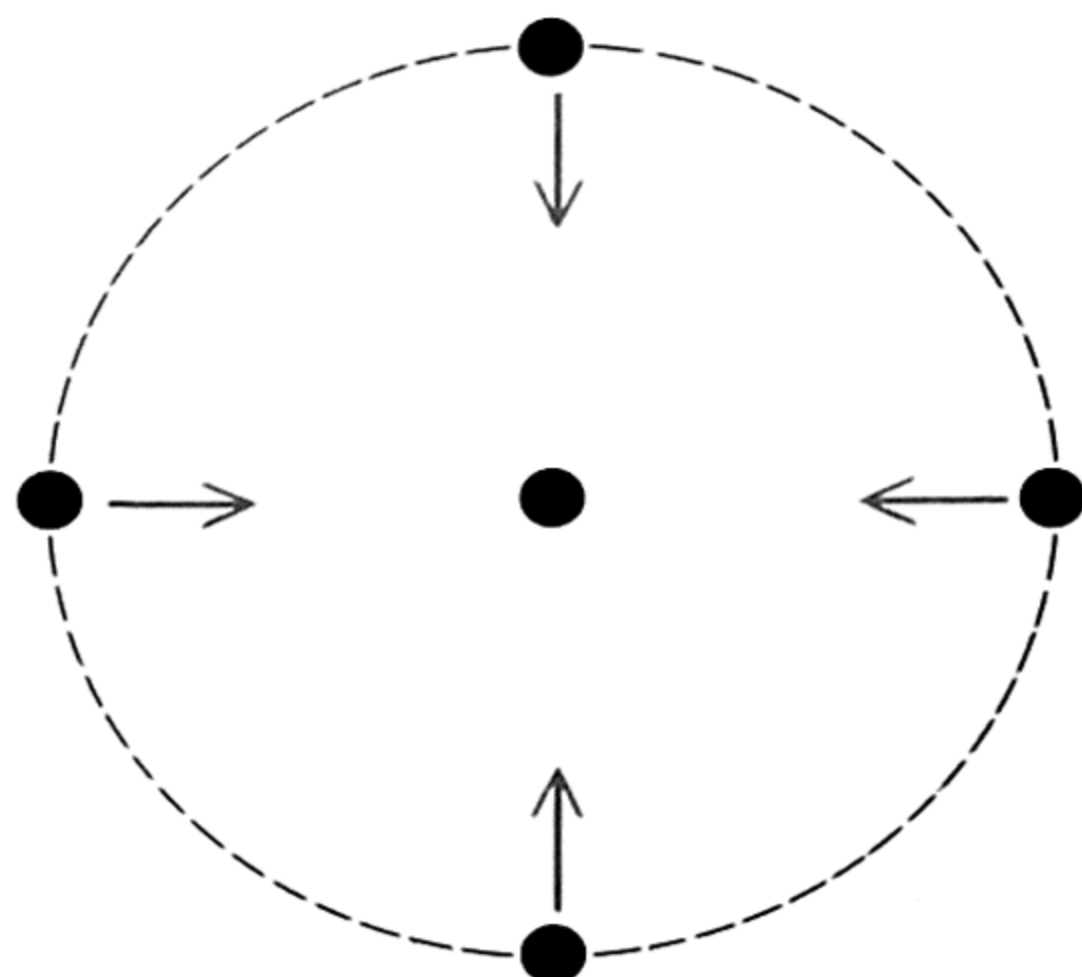
All the characteristics of gravity that have been listed apply equally well to electrostatic forces. These too are central and obey the inverse-square law. They differ only in being very much stronger than gravity, and in having two possible directions: attractive and repulsive.

When we consider magnetic forces, however, we find a different situation. If two magnets are held near each other, the force between them depends not only on their distance, but also on their relative direction with respect to their north-south axes [*see illustration at right on next page*]. Because of its mathematical properties, a force that behaves this way is known as a tensor force.

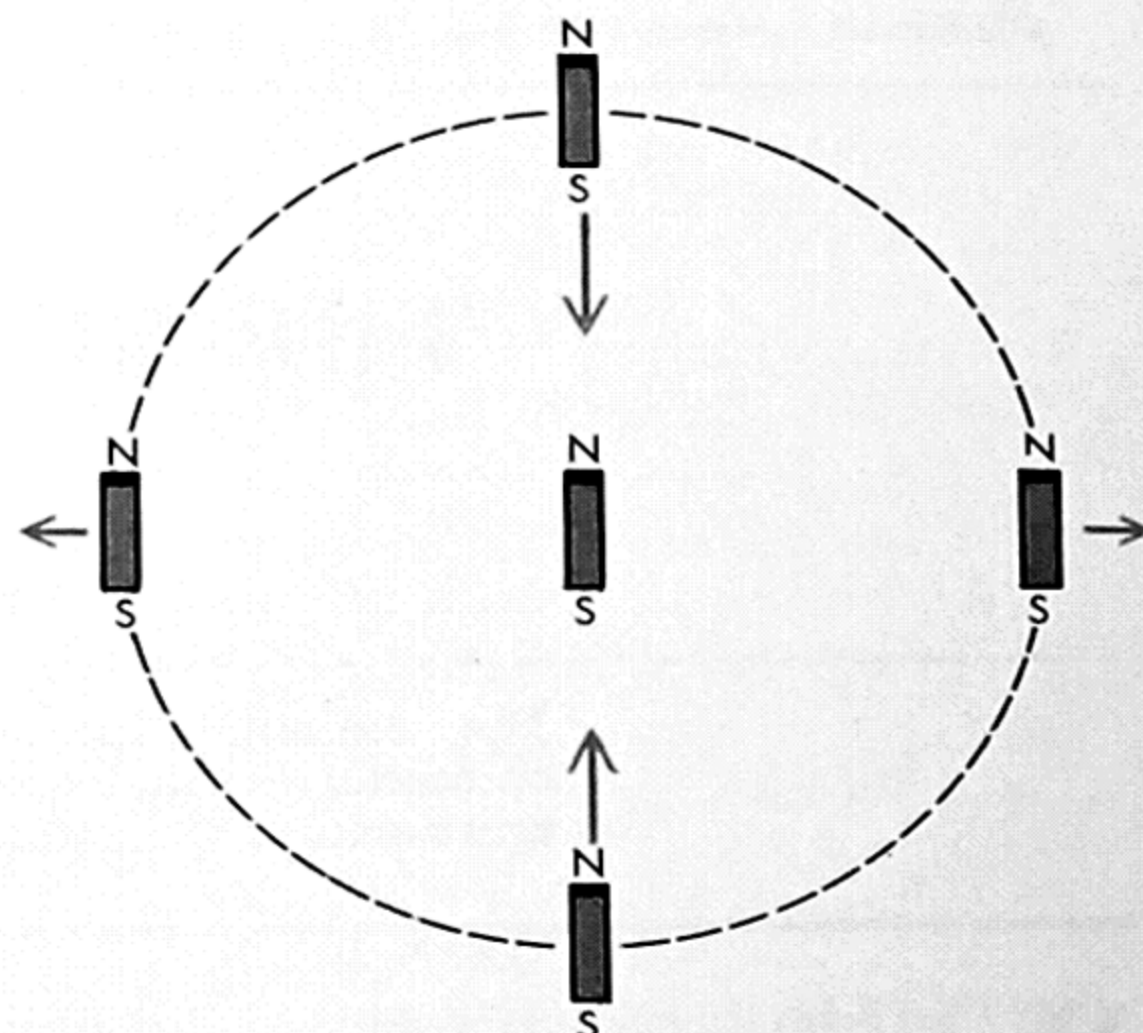
To find the force between a pair of magnets experimentally would be somewhat more difficult than to determine the nature of gravitational or electrostatic force. Many more measurements would be necessary in order to exhaust the various possible relative directions. In practice, of course, no measurements are necessary. We know enough about magnetism to compute the force between any pair of magnets in any relative position. Not only that, but magnetic, electric and gravitational forces can all be deduced from underlying theories.

The Nuclear Force

Now let us turn to the nuclear force. When the problem first presented itself, about all that could be said was that the force of attraction between particles in the nucleus must be extremely strong, but can only extend over a very short range. The fact that positively charged particles (protons) are bound together in nuclei showed that the nuclear attrac-



GRAVITY AND MAGNETISM are contrasted in these diagrams. Gravitational force (*left*) is "central"; i.e., it depends only on distance between attracting masses (*dots*) and not on their rela-



tive direction. Magnetic force (*right*) is a "tensor" force, differing for different directions with respect to magnetic axes. Arrows represent forces on outer bodies due to presence of body at center.

tion was stronger than the force of electrical repulsion. We now know that at a distance of one fermi (10^{-13} centimeter) the nuclear force is 35 times as strong as the electrostatic force and 10^{38} times stronger than gravity. (The fermi is the natural unit of length in this domain of physics. A nucleus measures a few fermis across.) At somewhat shorter distances it is even stronger. And at still shorter distances it reverses its direction and becomes repulsive. On the other hand, at distances beyond a few fermis the nuclear force rapidly drops to zero. The interaction of nuclei with particles a little way outside of them is almost wholly electric and magnetic.

When the English physicist James Chadwick discovered the neutron in 1932, it became clear that there are two types of nuclear building-blocks, or nucleons: the neutron and the proton. This meant that there might well be three kinds of nuclear force: proton-proton, neutron-neutron and proton-neutron.

How could they be measured? To use nuclei containing many protons and neutrons would only be compounding the difficulties. It is very hard to calculate the behavior of an assembly of many particles even when the law of force between them is known. To try deducing a law of force, or possibly several different laws, from the properties of the assembly is hopeless. The only practical way to begin was by studying isolated pairs of nucleons.

Nature has provided just one such

combination—the deuteron. This nucleus of heavy hydrogen consists of one proton and one neutron. More than 20 years ago studies of the behavior of deuterons began to reveal some of the details of the nuclear force. It at once turned out to be more complicated, as well as more powerful, than anything that had been known previously. The attraction between neutron and proton proved to be of at least *two* kinds. In part it was central: the same for all relative directions of the particles, like gravity. In part it was tensor: differing for different directions, like the force between two magnets.

Although we could not have predicted in advance that there would be a noncentral component, we did know that there might be. In order to have a force that depends on relative direction there must be some standard of reference by which one direction can be distinguished from another. In the case of electrostatic forces, for example, there is no such standard. A charged sphere looks exactly the same from any angle. Hence electrostatic forces cannot possibly be anything but central. A magnet, on the other hand, has a north-south axis. Seen from different vantage points, the axis takes on different orientations and thereby distinguishes among them. Thus magnetic force can be, and in fact is, noncentral.

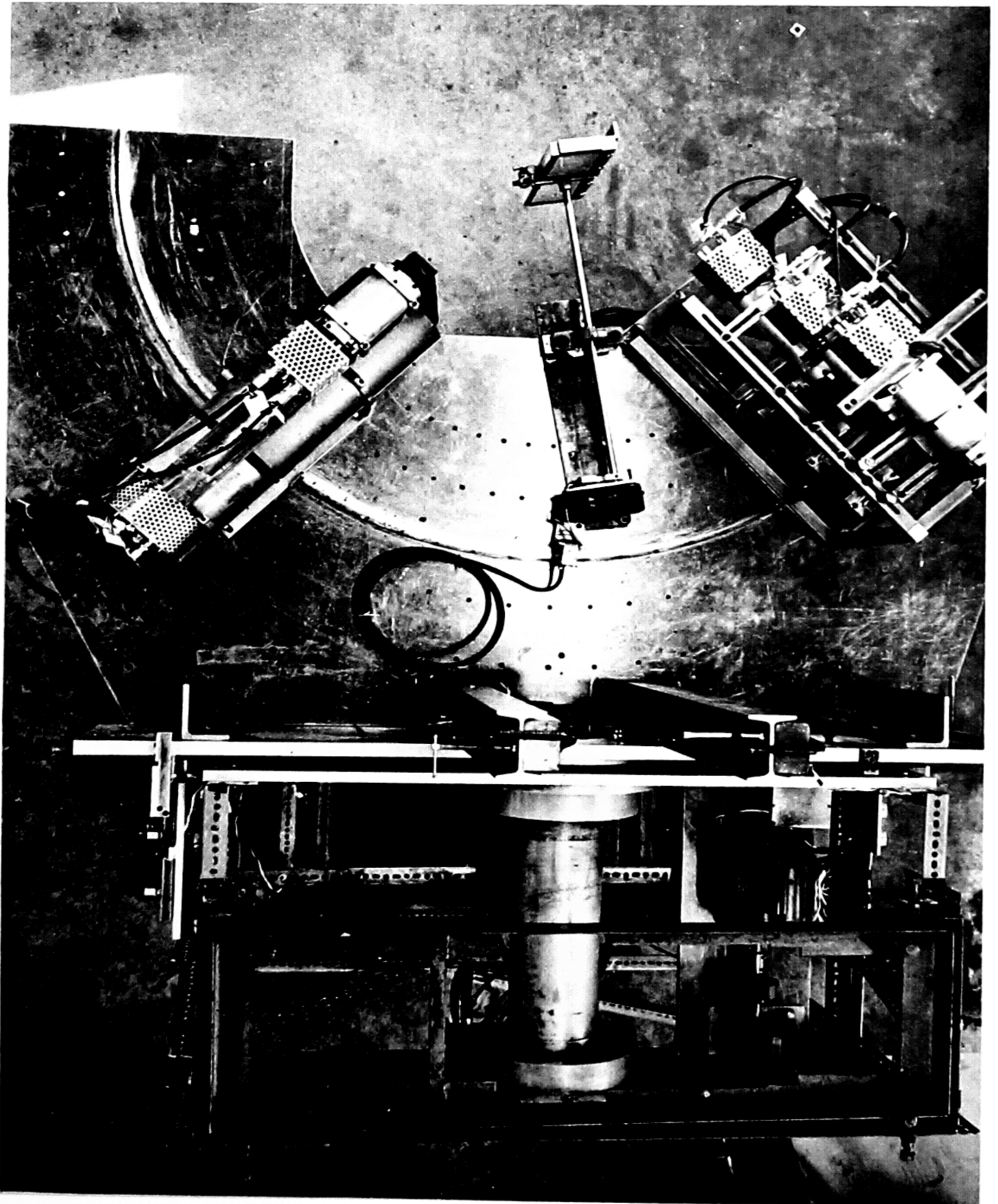
Nucleons also have a built-in direction indicator: the axis about which each one eternally spins. Hence they too can interact differently in different relative posi-

tions. (By convention the spin axes are assigned a direction depending on the sense of the rotation around them. If you curl the fingers of your right hand in the direction of the spin, your thumb gives the direction of the spin axis.) Notice that the existence of a reference direction does not guarantee that there will be a noncentral force. The sun and the planets have spin axes, but nature has not chosen to make gravity a noncentral force.

The properties of the deuteron, then, demonstrated that part of the proton-neutron force is noncentral and began to yield some specific figures for the strength of the force. But only a limited amount of information can be obtained from the deuteron. It tells only about the force between neutron and proton and only at the rather large distance of several fermis. The nucleons in the deuteron are much less tightly bound than the nucleons in heavier nuclei. Furthermore, the spin axes of the two particles always point in the same direction and can give information only about this relative orientation.

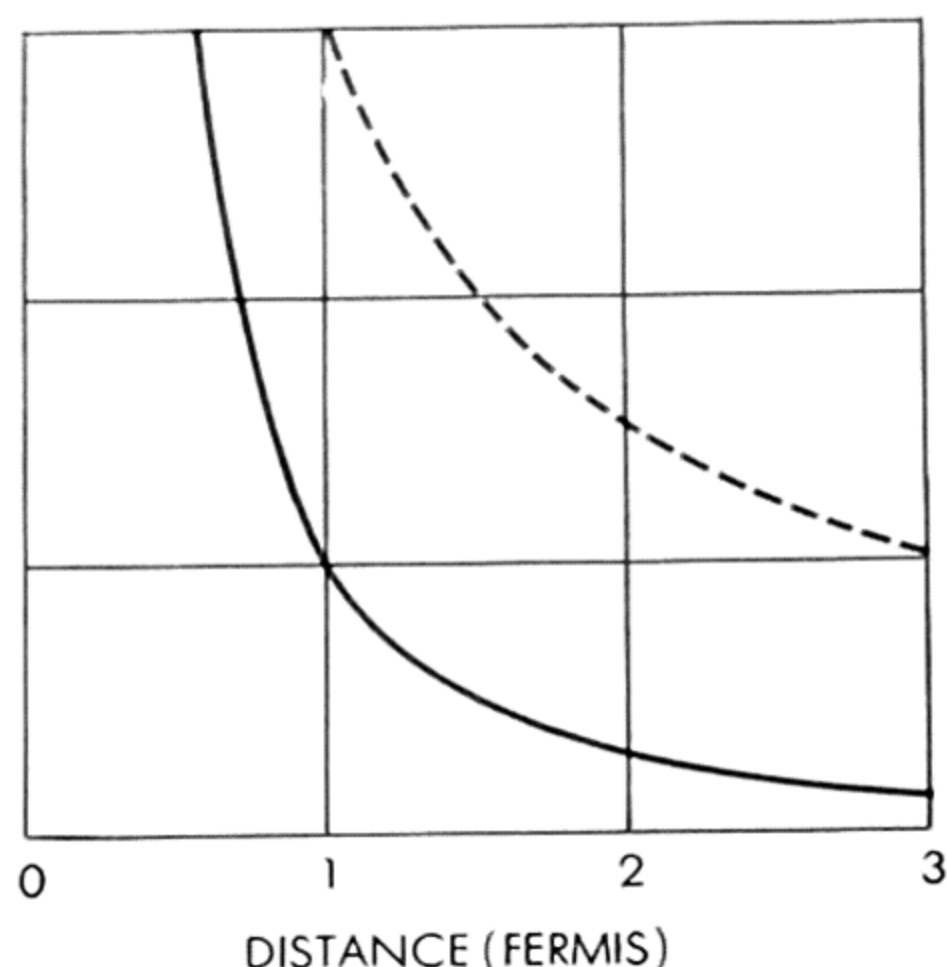
Scattering

To learn more we must make different kinds of two-nucleon systems. We do this by shooting one nucleon toward another at high speed. For a brief moment they come within range of the nuclear force, then separate again. In the process the projectile nucleon swerves from its



DOUBLE-SCATTERING EXPERIMENT is photographed at the University of Rochester. The first scattering target is located within the accelerator (*not shown*). Second target appears at top center.

Just below it are a neutron counter (*left*) and a proton counter (*right*) which respectively record the recoiling neutrons from a deuteron target and the deflected protons of the incident beam.



INVERSE-SQUARE FORCE, such as gravity, is represented by solid curve; corresponding potential energy, by broken curve.

original path and the target nucleon is pulled out of position. By studying the deflections we get an idea of the force that caused them.

Of course we do not actually deal with a single pair at a time. Instead we send a dense beam of nucleons from a cyclotron or other accelerating machine through a target material and count the numbers of particles emerging at various angles. The beam may consist of protons or neutrons. The target may be hydrogen, whose nucleus is a single proton, or deuterium, whose nucleus has a proton and a neutron, or in some cases a heavier element. Thus all the combinations can be studied: proton-proton, neutron-neutron and neutron-proton. By using beams of different energies we obtain information about the force at different distances, as we shall see.

In describing the process by which a beam is scattered it is usual to speak of "collisions" between incident and target particles. But a better picture is a near miss, with incident and target particles orbiting around each other for the brief period that they are within effective range of the nuclear force. Scattering experiments are analogous to the observations on the orbits of the planets, which led to Newton's law of gravitation. They differ, however, in that the orbits are not closed ellipses, but open curves like the path of a rocket vehicle that swings once around the moon and then heads out into space. The similarity would be greater if the vehicle were as

heavy as the moon and pulled the moon off course as it went by.

Even then the analogy would be imperfect. Any picture of an atomic or nuclear process that involves distinct little balls speeding along well-defined paths is misleading. Not only are the particles far too small to see and to trace, but also the uncertainty principle of quantum physics tells us that sharp trajectories do not even exist except at energies much higher than those we shall deal with. The only rigorous way to predict, or even to describe, the interactions we are interested in is to think of the particles as waves, and of their mutual effects as a mingling of waves. In our discussions and diagrams we use both waves and particles. This is only to convey a rough, intuitive idea of what is going on. The actual computations whose results we are describing are carried out by the consistent but abstract method of wave mechanics.

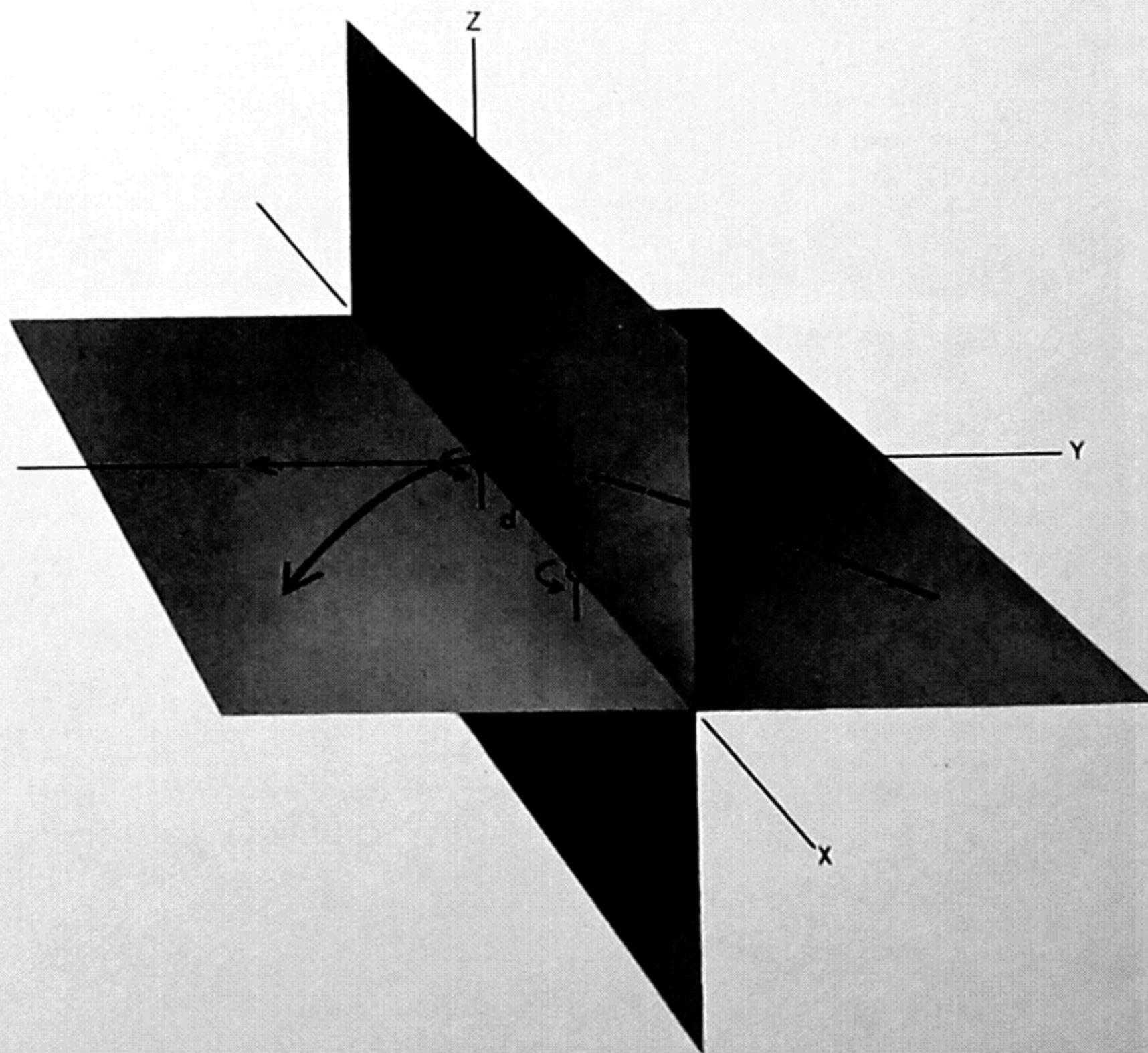
Angular Momentum

With these reservations in mind we can return to semifictional particle lan-

guage to describe a scattering event. To simplify matters we may suppose for the moment that the target nucleon remains stationary while the bombarding, or incident, nucleon swings around it in a roughly hyperbolic orbit [see illustration below]. When the force is attractive, the orbit curves toward the target; when repulsive, the orbit curves away.

As has already been mentioned, both incident and target nucleons are spinning, and the relative direction of their spin axes may have a great effect on the force between them. If they were "classical" particles, the spins could be in any relative direction whatever. Quantum mechanics makes things much simpler; a pair of nucleons can spin only in the same direction or in opposite directions. In other words their spin axes must be either parallel or antiparallel.

This rule applies to any particle which, like a nucleon, has a spin of one-half unit. The unit is Planck's constant, h , divided by 2π , and it measures the angular momentum associated with any type of rotational motion in the quantum domain. Thus a nucleon has a spin angular-momentum of one-half unit. If the



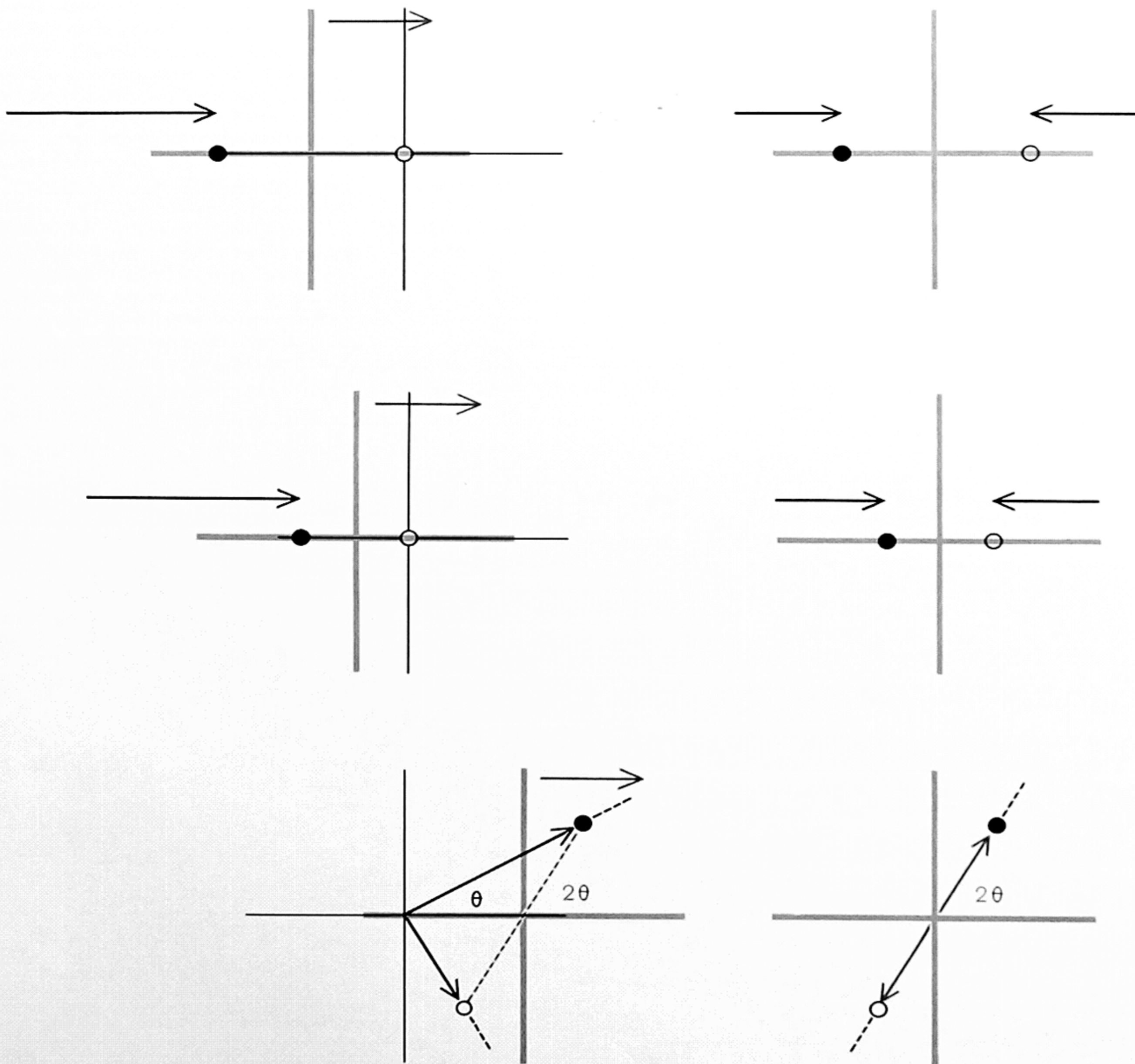
SCATTERING EVENT is illustrated in this greatly simplified drawing. Projectile nucleon (solid dot) follows curved trajectory (colored line) around target nucleon (open dot). (In fact both particles move.) Small curved arrows represent spins; small vertical arrows, the spin angular-momentum vector; horizontal arrow, linear-momentum vector; large vertical arrow, the orbital angular-momentum vector; d , the distance of closest approach.

spin axis is in one direction, say up, the angular momentum is considered to be plus one-half unit. If it is in the opposite direction, or down, the angular momentum is minus one-half unit.

A system of two interacting nucleons has a total spin angular-momentum,

found by adding the spins of the particles. When the individual spins of the two nucleons are parallel, the total spin is one; when they are antiparallel, the total spin is zero. Hence for the case of parallel spins there is a net total spin, providing a reference line with respect

to which relative directions between the two nucleons can be differentiated. Two nucleons with parallel spins can, and in fact do, have a noncentral component in their interaction. But for antiparallel spins the total spin is truly zero. It is as if the separate particles were not spin-



CENTER-OF-MASS coordinate system (*colored axes*) is illustrated in relation to laboratory coordinate system (*black axes*) at left, and as a stationary frame of reference at right. Incident particle is shown as solid dot; target particle, as open dot. Black arrows

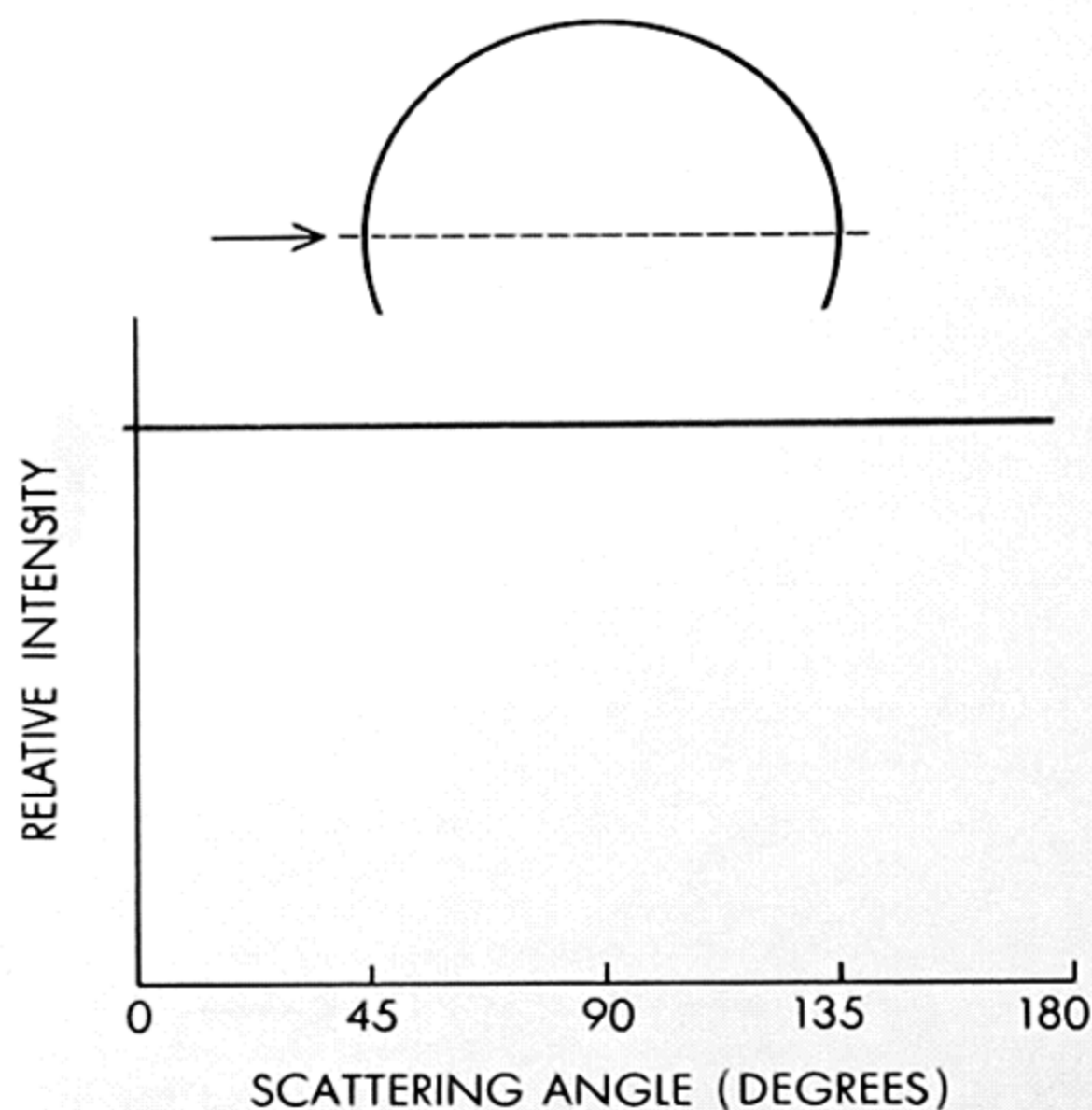
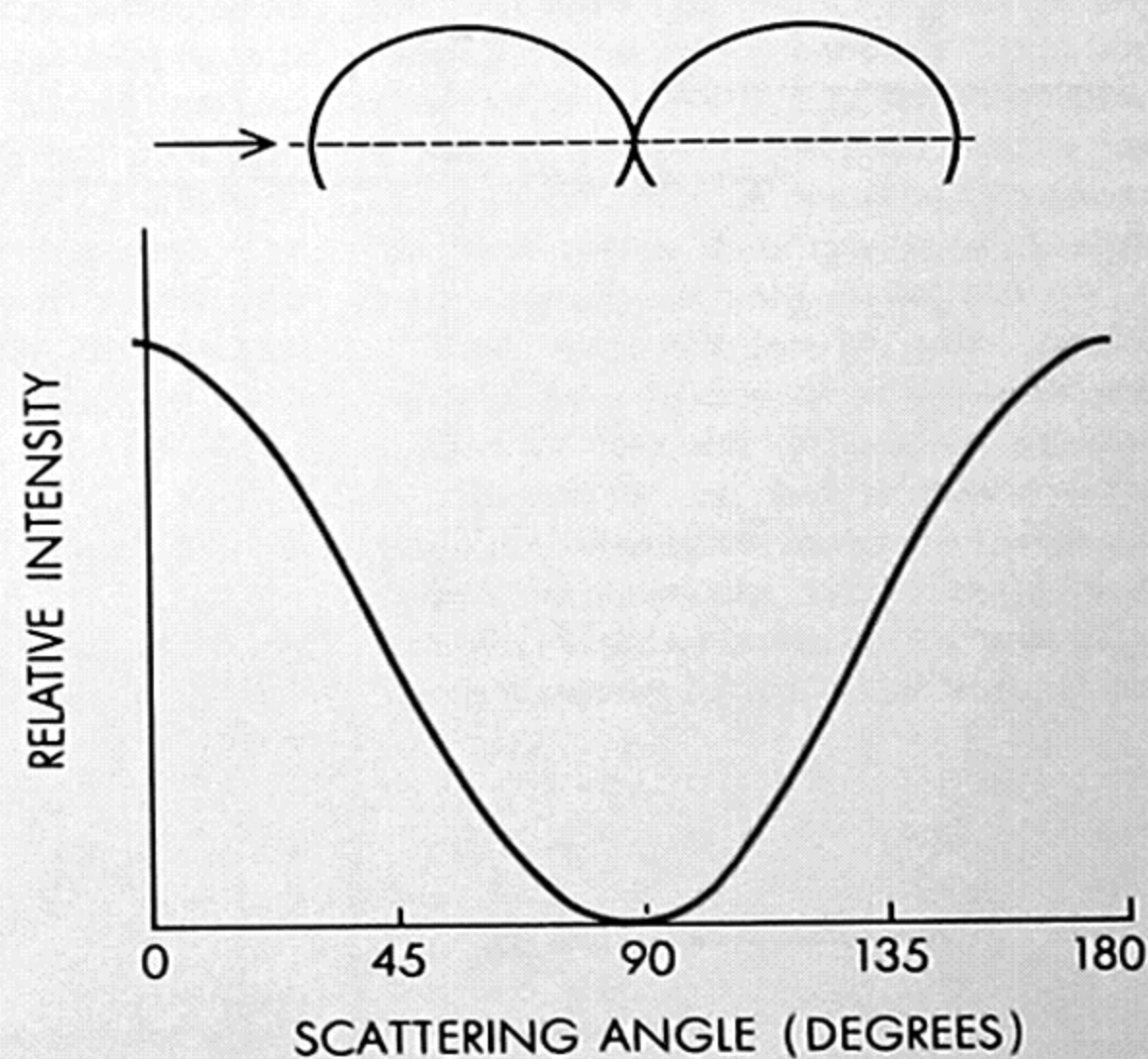
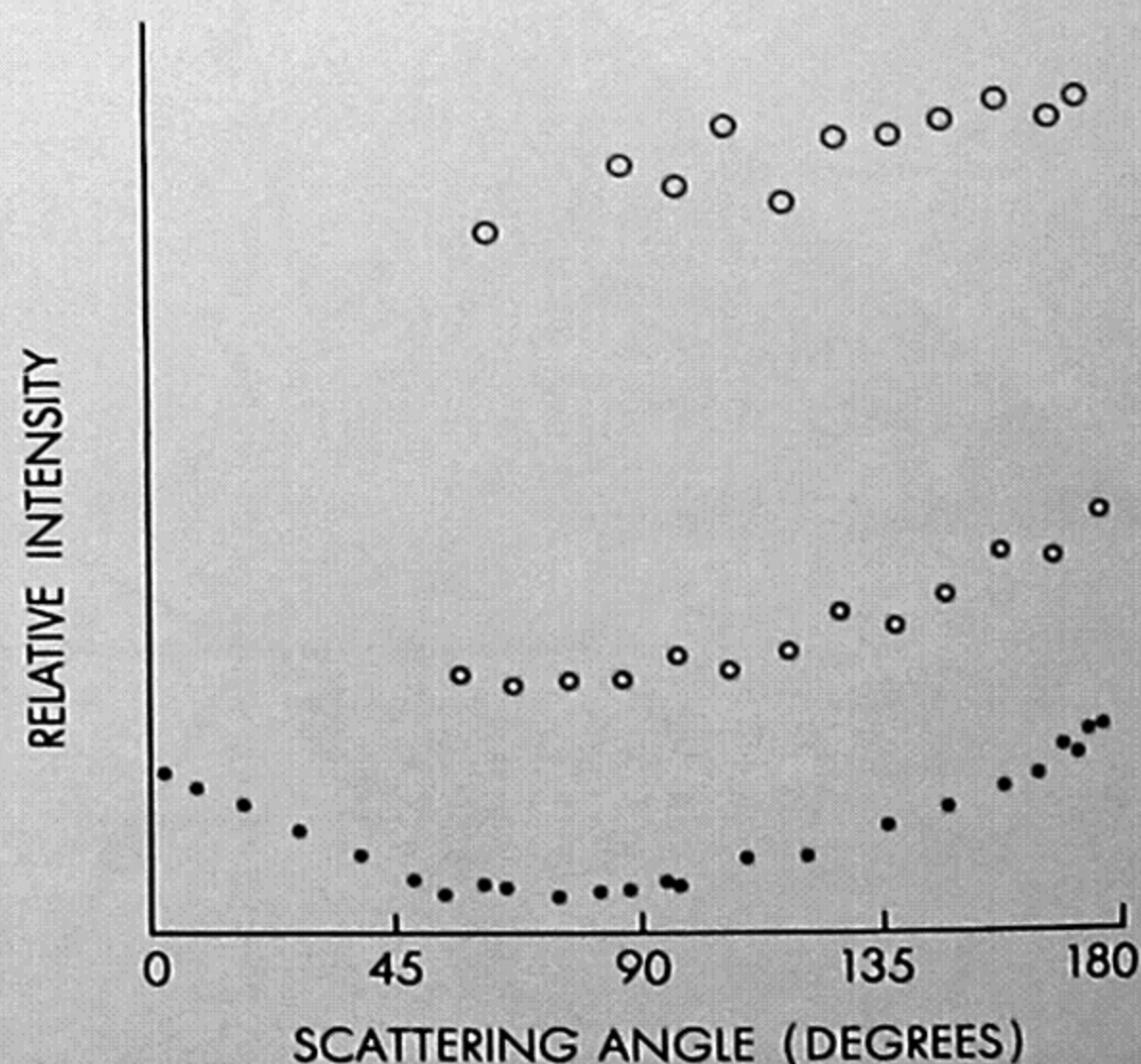
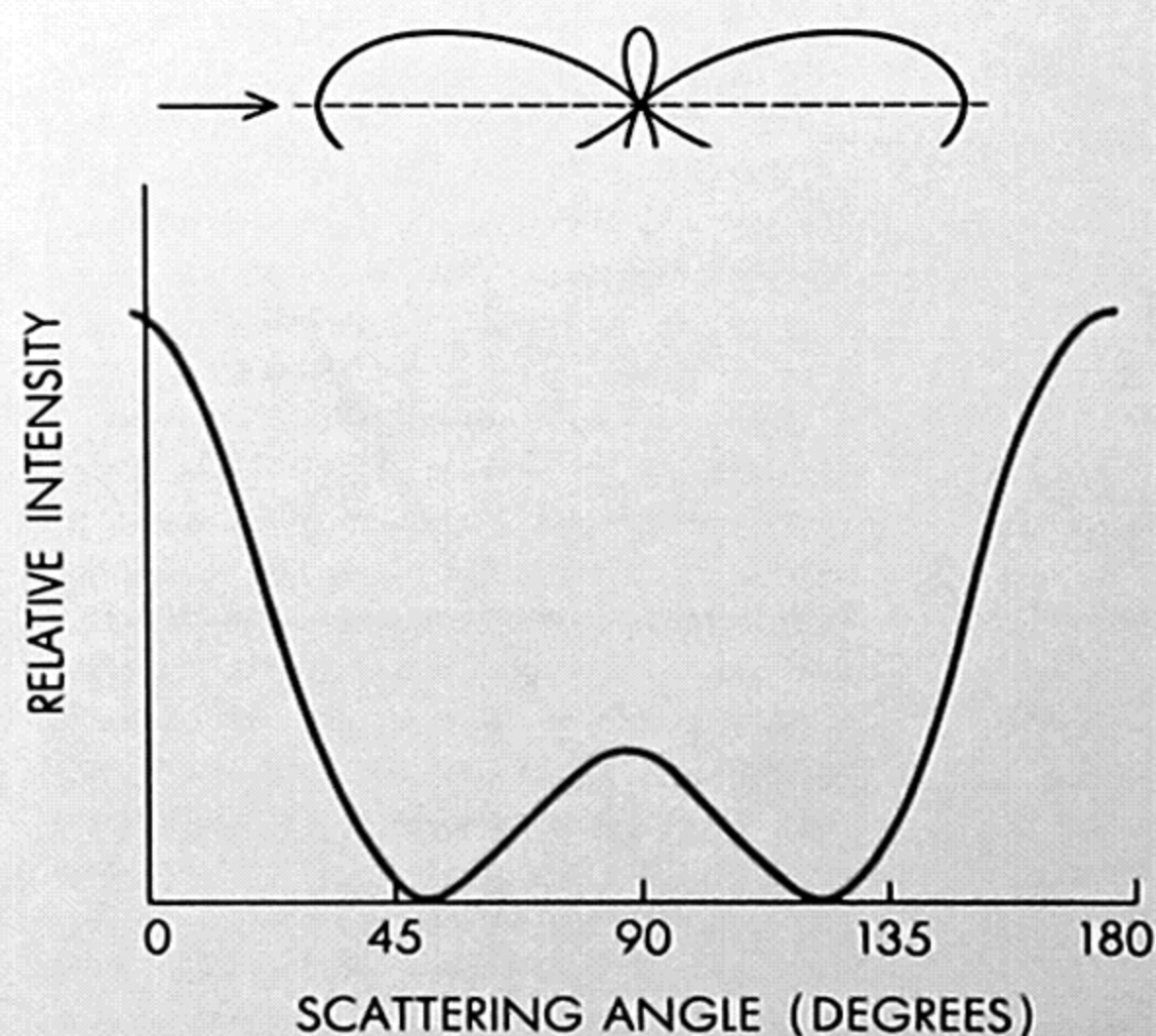
indicate velocities in laboratory frame; colored arrows, in center-of-mass frame. Views of approaching particles appear at top and center and a view after collision at bottom. Scattering angle, theta (θ), in the laboratory system is half that in center-of-mass system.

ning at all. There is no reference direction and hence no noncentral component; the force between the two nucleons is entirely central.

In addition to the spins of the separate particles the scattering system has another rotational feature: the curved

orbits of the particles. Associated with this type of motion is an orbital angular-momentum. In the simplest case, of uniform motion in a circle, the amount of orbital angular-momentum is found by multiplying the linear momentum by the radius of the circle. For the open curve

of a scattering orbit we can find the orbital angular-momentum by multiplying the linear momentum of the incident particle at the moment it is nearest the target by this distance of closest approach. As with spin angular-momentum, the direction of orbital angular-

S WAVE ($l=0$)P WAVE ($l=1$)D WAVE ($l=2$)

SCATTERING PATTERNS for pure S-wave, P-wave and D-wave cases, corresponding to the orbital angular-momentum numbers (l) zero, one and two, are plotted at top left, top right and bottom left respectively. Points at lower right are experimental values of

relative particle counts in neutron-proton scattering experiments at neutron energies of 18 million electron volts (*top*), 40 Mev (*middle*) and 150 Mev (*bottom*). The geometrical designs appearing above the S, P and D curves are polar plots from zero to 180 degrees.

momentum is represented by an arrow pointing perpendicularly to the plane of the curved path in accordance with the right-hand rule.

In the quantum world, angular momentum is a particularly important quantity, and it is subject to two crucial conditions. First, when the two nucleons have a spin of one (that is, when their individual spins are parallel), the total-spin angular-momentum vector can have three, and only three, directions with respect to the orbital angular-momentum vector: the two can be parallel, perpendicular or antiparallel. This relationship provides a criterion for distinguishing a second type of noncentral force. In addition to the relative direction between the two spinning particles, which is the basis of the tensor type of noncentral force, there is now the relation between total-spin angular-momentum and orbital angular-momentum. Nature has taken advantage of this opportunity to differentiate and set up a second component of noncentral force between two nucleons which we call the spin-orbit force. (Like the tensor force it can exist only when the spins of the particles are parallel;

when they are antiparallel, there is no spin angular-momentum and no spin-orbit force.)

The second quantum characteristic of orbital angular-momentum is that it can have only integral values. It can be zero, one unit, two units and so on. What does zero orbital angular-momentum mean? In classical terms it means that the distance of closest approach of the two nucleons is zero. We must think of the incident particle passing directly through the target particle. This is rather hard to imagine, so we must here have recourse to wave ideas. It is easy enough to picture one wave passing through another. For the higher values it is possible to picture curved paths, but they "really" represent wave processes too. The wavelengths of these waves depend on the orbital angular-momentum and on the distance of closest approach, as we shall see.

Beam Energy

So much, then, for the complicated anatomy of a single scattering event. Imagine now a stream of billions of par-

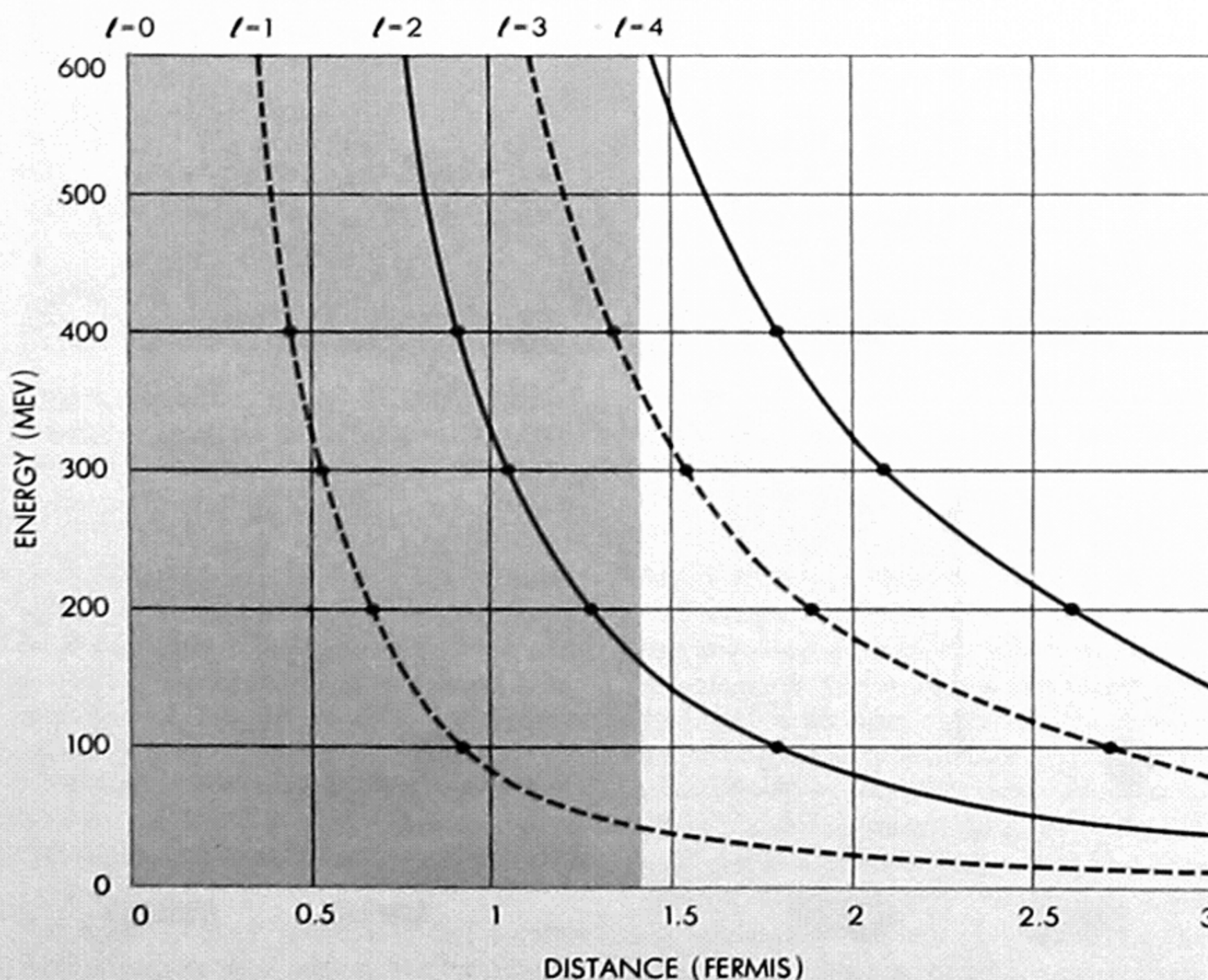
ticles approaching a still larger number of target nuclei, as in an actual scattering experiment. First let us make the unlikely assumption that each incident particle is headed dead center at a target. In other words all the orbital angular-momenta are zero. If the particles can be thought of as passing through each other, we might suppose that the beam would simply continue straight ahead and that no sideward scattering would be observed. But not at all. If detecting counters are moved around the target, particles are found in varying amounts in all directions from zero deflection through 90 degrees (for a hydrogen target). The effect can be explained in wave terms as a diffraction process, like the angular spreading of a beam of light when it passes through a transparent diffraction grating.

Next assume that all the incident particles are aimed in such a way as to give orbits with an angular momentum of one unit. Now a quantum condition comes into play. As we have said, orbital angular-momentum is the linear momentum (mass times velocity, or mv), multiplied by the distance of closest approach, which we shall call d . The product $mv d$ must be equal to one unit in the case we are considering. Thus the greater the value of v , the smaller d will be, which means the nearer the approach to the target. It is customary to describe particles in a beam in terms of their energy, which of course depends on velocity, rather than in terms of velocity itself. So we can equally well say that the greater the energy, the nearer the approach. Conversely, the less the speed or energy, the greater d must be, or the farther the particles are separated at their closest approach.

But nuclear forces extend over very short distances. If the speed of the incoming particle is too low, making d too large, the projectile and target will remain entirely out of range so far as the nuclear force is concerned. Then there will be no scattering at all. In order to get scattering at an orbital angular-momentum of one unit, the energy of the incident beam must exceed a certain minimum.

The same is true for the higher orbital angular-momenta, the minimum effective energy being successively greater. (If $mv d$ equals two units, v must be still larger to make d small enough to fall within range of the nuclear force.)

If it were possible to have a pure beam of orbital angular-momentum one, then if the energy were above the necessary minimum, an angular scattering pattern



ENERGIES associated with orbits of various angular momenta (l) are plotted as curved lines, with even values solid and odd values broken. Curve corresponding to zero momentum is the coordinate axes themselves. Colored horizontal lines represent energies of particle beams, and their intersections with the curves show the corresponding distances of closest approach. Shaded region within 1.4 fermis represents the range of nuclear force.

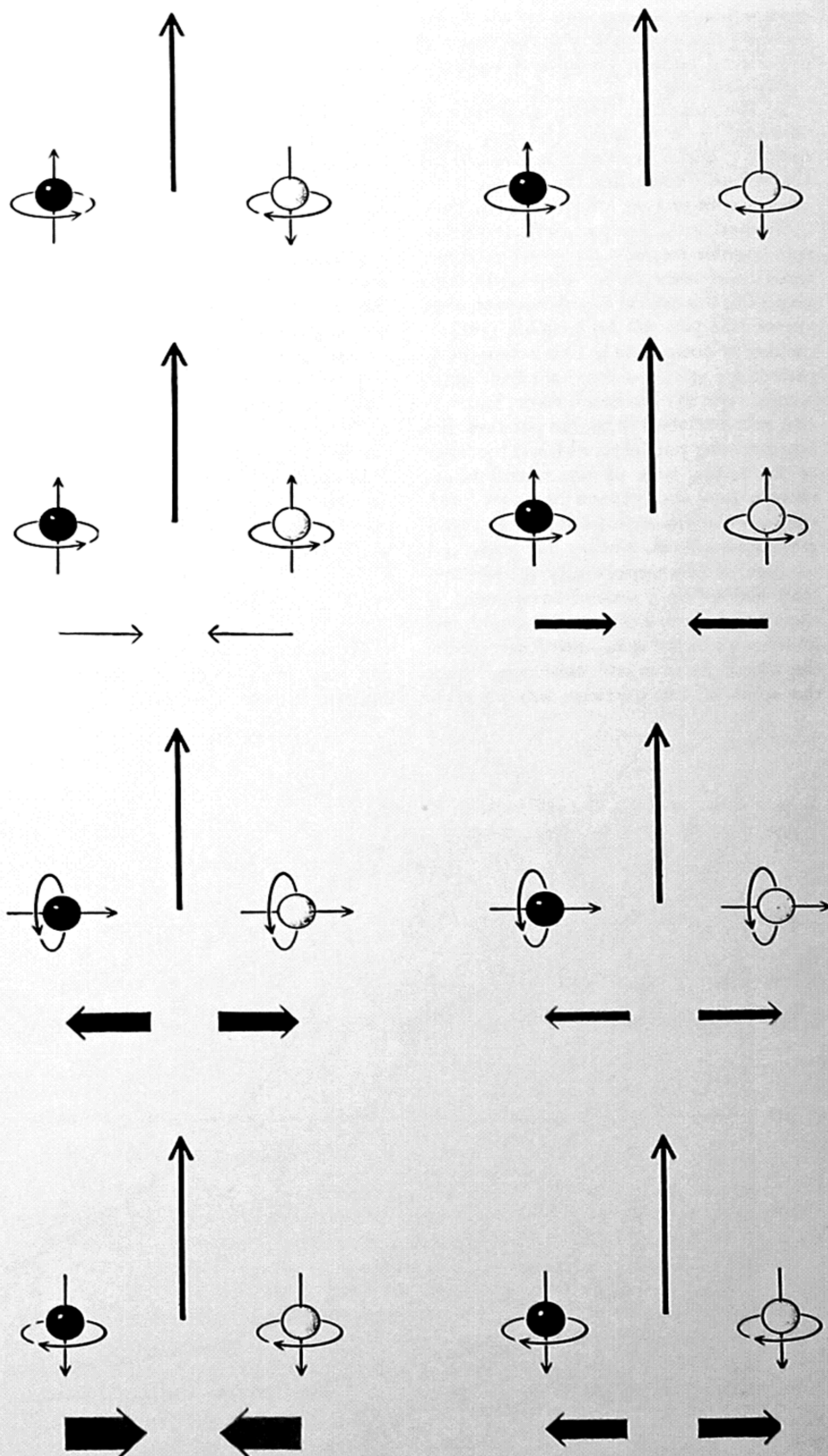
would be observed that is quite different from the one for orbital angular-momentum zero. Similarly, the values two, three and so on would each give their own distinctive pattern.

Patterns of Scattering

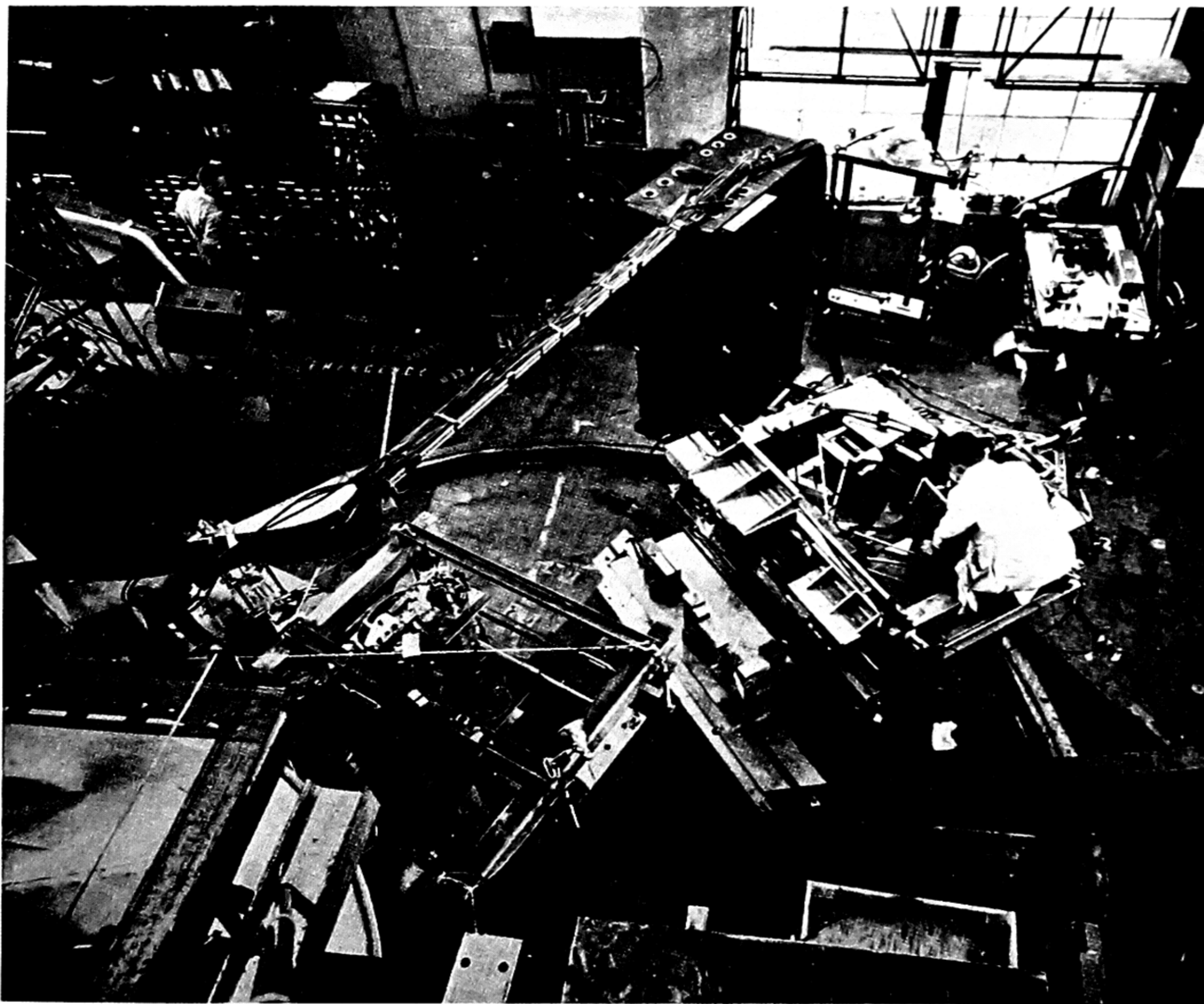
What do the scattering patterns look like? The raw angular distributions as they are obtained in the laboratory have rather complicated shapes. But they become much simpler if translated into the so-called center-of-mass coordinate system. This reference system, universally employed in collision problems because of its mathematical convenience, has as its zero point the center of mass of the two particles. Where the two have equal masses, as they do in all the experiments we are considering, this point is midway between the two particles at every instant [see illustration on page 571]. If we imagine ourselves riding on the center of mass and consider ourselves at rest, then before the collision the particles would seem to be moving toward each other, each one having half the speed of the moving particle in the laboratory frame of reference. After the collision the particles will seem to be moving in exactly opposite directions, as can be seen in the illustration, and the scattering angle will be twice that measured in the laboratory frame of reference.

In the center-of-mass system the scattering pattern for zero orbital angular-momentum is spherical. An equal number of particles is deflected at every angle. For an orbital angular-momentum of one unit the pattern becomes a solid figure eight; the number of scattered particles is greatest in the forward direction, decreases to zero at 90 degrees, then increases again from 90 to 180 degrees and repeats on the other side of the center line, from 180 to 360 degrees [see illustration on page 572]. Higher orbital angular-momenta yield more complicated patterns, with several lobes. These shapes are commonly denoted by the letters S, P, D, F, G, H, etc., for patterns corresponding to the values 0, 1, 2, 3, 4, 5, etc., of orbital angular-momentum.

Although the curves are obtained by counting particles scattered at different angles with respect to the direction of the incident beam, they may be thought of in another way. They are pictures of the different waves that can make up a two-particle system. We speak of S-wave, P-wave, D-wave scattering and so on. The wavelength of all the waves grows shorter as the energy of the inci-



NONCENTRAL FORCES between pairs of nucleons depend on the relative orientation of the particles. The tensor force is illustrated in the left-hand column; the spin-orbit force, in the right-hand column. The thickness of horizontal arrows is proportional to the size of the force in each case. The small, straight arrows passing through each particle are spin angular-momentum vectors. The large vertical arrows are orbital angular-momentum vectors.



TRIPLE-SCATTERING EXPERIMENT at the University of Rochester employs three targets. The first is within the synchrocyclotron, the edge of which appears at lower left. Polarized beam

passes through cylinder of liquid hydrogen (*lower left center*) and then to third target-assembly mounted on track at right. This assembly can move around track to count particles at different angles.

dent beam increases (varying inversely as the square root of the energy). It can be shown that the distance of closest approach is equal to the wavelength times number of orbital angular-momentum units.

In practice it is possible to have pure S-wave scattering when the energy of the incident beam is low. We do not see pure P, D or higher waves, however. As the beam energy increases, the various higher modes are added to the S-wave pattern in greater and greater degree. This means that the beam contains particles of various angular momenta, with the higher values contributing increasingly to the scattering pattern as the

beam energy increases.

Perhaps the easiest way to visualize the situation is to plot a series of curves, where each curve shows the energy associated with one value of the orbital angular-momentum number (usually denoted by l) at various distances of separation [see illustration on page 573]. On this graph a straight horizontal line represents beam energy in a scattering experiment. Any such line will cross the successively higher orbital angular-momentum curves at greater and greater distances, which represent the increasing distances of closest approach. If we indicate the effective range of the nuclear force on the same graph, we can

see at a glance which angular momenta will contribute to scattering at a given energy.

The "range" of the nuclear force is generally taken as 1.4 fermis. At this distance it does not drop to zero, but to about one third of its maximum value. In about two more removes of 1.4 fermis, the force becomes essentially zero. Any momentum curve crossing the beam-energy line beyond 1.4 fermis does not contribute very strongly to the scattering pattern.

Resolving Power

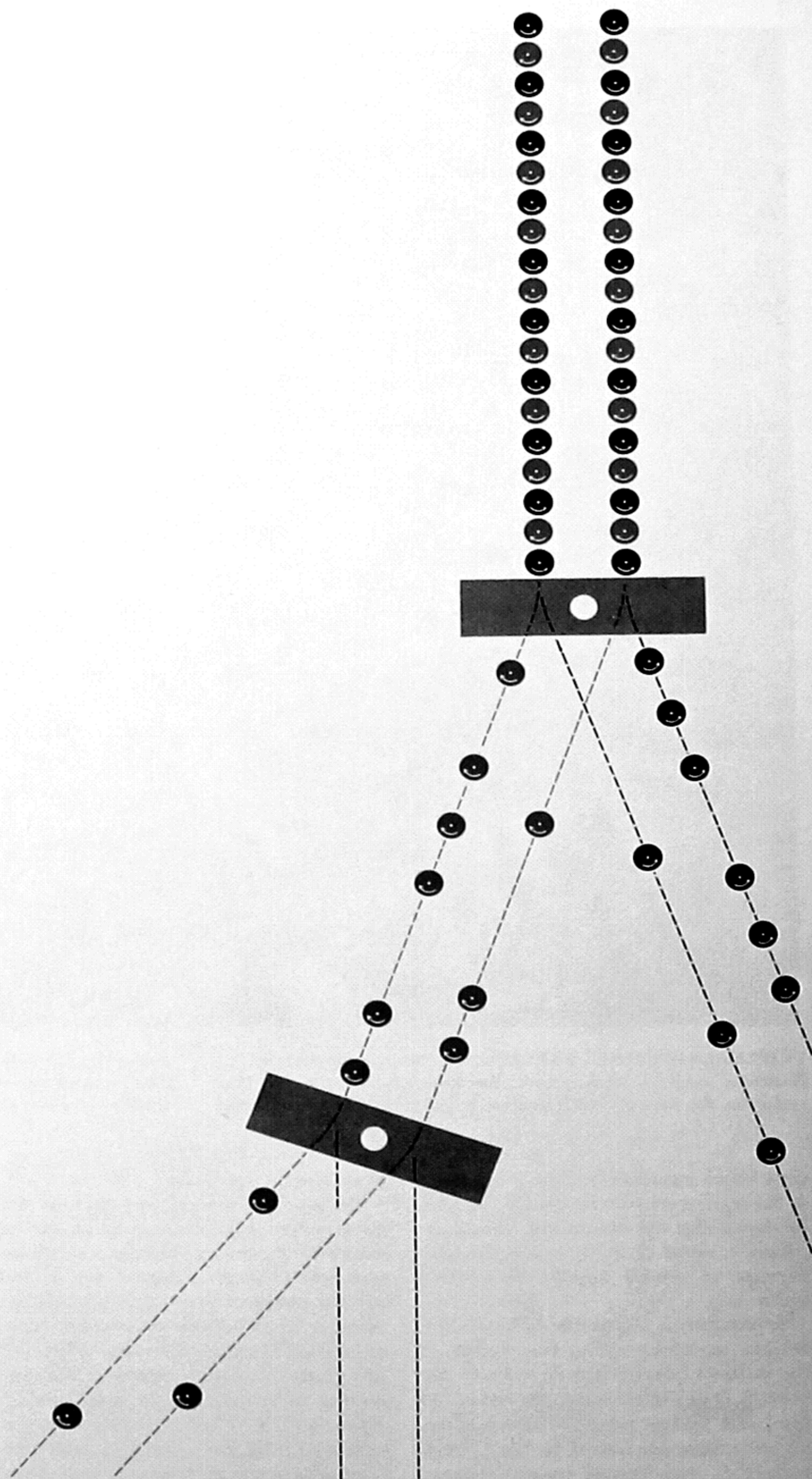
What the graph demonstrates is the

increasing resolving power of higher and higher beam-energies. As energy increases, waves of higher orbital angular-momentum enter into the scattering process. As we have seen, the distance of closest approach of the two nucleons is equal to the wavelength times l , so that larger l means shorter wavelength. In addition, for any value of l , higher energy means shorter wavelength. Thus all the waves—S, P, D, etc.—have shorter wavelengths at higher energies. Just as in an optical observation, the shorter the wavelength compared to the dimensions of the object being studied, the finer the detail that can be seen. Here the “object” is a pair of particles within the range of the nuclear force: about 1.4 fermis. The more wavelengths that fit into this range, the higher the resolving power.

At first glance it might seem that resolving power could be increased indefinitely, up to the highest energies available in accelerating machines. However, above an energy of 300 million electron volts (300 Mev) a new factor complicates the scattering process. At such energies the nucleons do not simply orbit around one another and separate; instead new particles are created in the collision. These are pi mesons, or pions, which are the “field quanta” of nuclear forces [see “Pions,” by Robert E. Marshak; *SCIENTIFIC AMERICAN* Offprint 226]. The creation of additional particles greatly complicates the analysis of scattering.

Fortunately we need not consider energies above 300 Mev if we are interested in the force holding nuclei together, as opposed to the more general nucleon-nucleon interactions of particle physics. Within the nucleus the average kinetic energy of the constituent nucleons is about 25 Mev. The maximum energy ever acquired by an individual nucleon with respect to a neighbor is approximately 100 Mev. Scattering experiments at energies much higher than this do not represent conditions that could occur with any significant probability inside a nucleus. In this article we consider only experiments up to 300 Mev, thus limiting ourselves to the range of nuclear physics rather than of all particle physics.

Our picture of scattering is now complete except for one final quantum restriction: the famous Pauli exclusion principle. The principle says that no two identical particles of spin one-half unit can be in exactly the same quantum state. In the familiar case of atomic electrons orbiting around a nucleus, this



POLARIZATION experiment is diagrammed schematically. The original beam of nucleons (*top*) is unpolarized, with as many clockwise spins as counterclockwise. As particles pass nucleons in first target (*upper gray area*), they separate according to spin directions. Polarized beam passing through second target (*lower gray area*) scatters preferentially in one direction.

means that each orbit can contain only two electrons with spins in opposite directions. The application of the exclusion principle to the open "orbits" of the scattering process is somewhat different. Here the restriction is that pairs of (identical) nucleons with parallel spins must have odd orbital angular-momenta ($l = 1, 3, 5$, etc.). Nucleons with antiparallel spins can have only even orbital angular-momenta ($l = 0, 2, 4$, etc.). Hence for proton-proton or neutron-neutron scattering, two cases are to be distinguished: parallel spins, in which only odd scattering modes (P, F, H, etc.) enter, and antiparallel spins, which go with the even modes (S, D, G, etc.). In neutron-proton experiments the Pauli principle does not apply, and all the orbital angular-momenta, even and odd, contribute to scattering for both parallel and antiparallel spins.

Now throughout this detailed description of the dynamics of nucleon-nucleon interactions we have been speaking as if the actual particles were visible to us, and we could see them spinning, curving, diffracting and so on. It should be

remembered, however, that all we can really see are the recording devices of a particle-counter, or the dark spots on a photographic emulsion, representing relative intensities of a particle beam scattered in various directions. From these data the experimenter draws a curve, which he then presents to the theoretical physicist and asks: What kind of force produced such a graph?

The theoretician realizes that the patterns contain unknown mixtures of S-waves, P-waves, D-waves and so on; that they may correspond to antiparallel or parallel spins; and that the parallel spins may have pointed in different directions with respect to the orbital angular-momentum. And all these variations can lead to different forces. How can they be disentangled?

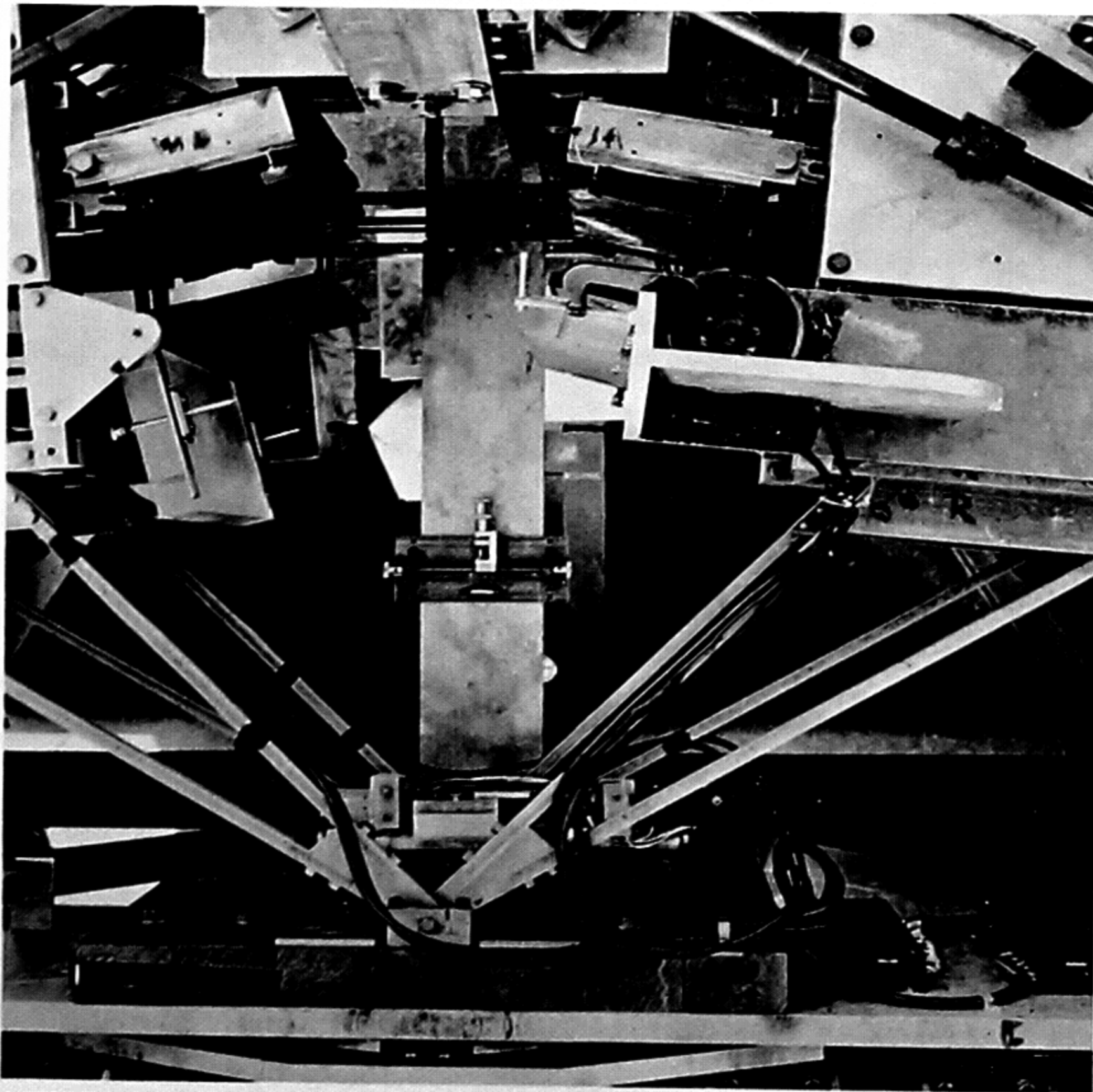
Polarization

For a long time the situation seemed almost hopelessly complicated. The earlier scattering experiments had helped outline the cruder features of nuclear forces, but the fine details were beyond

their power to illuminate. Then in 1953 C. L. Oxley and his colleagues at the University of Rochester, working with a 240-Mev synchrocyclotron, discovered that high-energy proton beams are strongly polarized after being scattered from a hydrogen target. The particles emerging at various angles contain a preponderance of one spin direction, up or down with respect to the plane of the incident and scattered beams. Here was a powerful new tool to help discriminate among the bewildering array of possible interactions. Soon a number of other laboratories took up polarization experiments along with Oxley's group. In particular the work was carried on at Harvard University and at the Atomic Energy Research Establishment in Harwell, England, with 150-Mev accelerators, and at the University of California and the University of Liverpool in England with 300-Mev machines. A number of theoretical physicists have applied themselves to analyzing the results with the help of large electronic computers. Out of this far-flung effort is emerging a much clearer picture of the nuclear force.

Looking back, we can see that polarization should have been expected to result from noncentral forces. To understand why, imagine a beam of unpolarized protons' (with as many upward spins as downward) passing to the right and left of target nuclei [see illustration on page 576]. Those passing to the right have orbital angular-momentum pointing up; those to the left, down. This means that on one side the upward spins are parallel to the angular momentum and the downward spins antiparallel; on the other side the situation is reversed. If the force is different for the parallel and antiparallel cases, the particles will be scattered differently depending on their upward or downward spin. Any difference in force is enough to upset the symmetry of scattering. Thus the spinning particles are segregated, one group going preferentially to the right and the other to the left. If the force changes from attractive to repulsive, as it can in some cases, the separation would be complete and the beams would be 100 per cent polarized.

The counters used to detect scattering, however, do not discriminate between the two types of spin. To gather the information contained in polarized beams an analyzer is needed, just as it is in polarized-light experiments. A second hydrogen target provides the necessary analyzer. If the particles approaching it are all, or almost all, spinning in the same direction, then they will be scat-



SCINTILLATION COUNTERS (rectangular plates to left and right of center at top) measure the numbers of particles scattered at a given angle to either side of the third target.

tered most strongly in the preferred direction. The scattering pattern will now show an asymmetry as between right and left. From this asymmetry the degree of polarization can be measured [see top illustration on page 579]. Thus double-scattering experiments yield much more information than single-scattering ones. They were the first to demonstrate unequivocally that the proton-proton force has noncentral components.

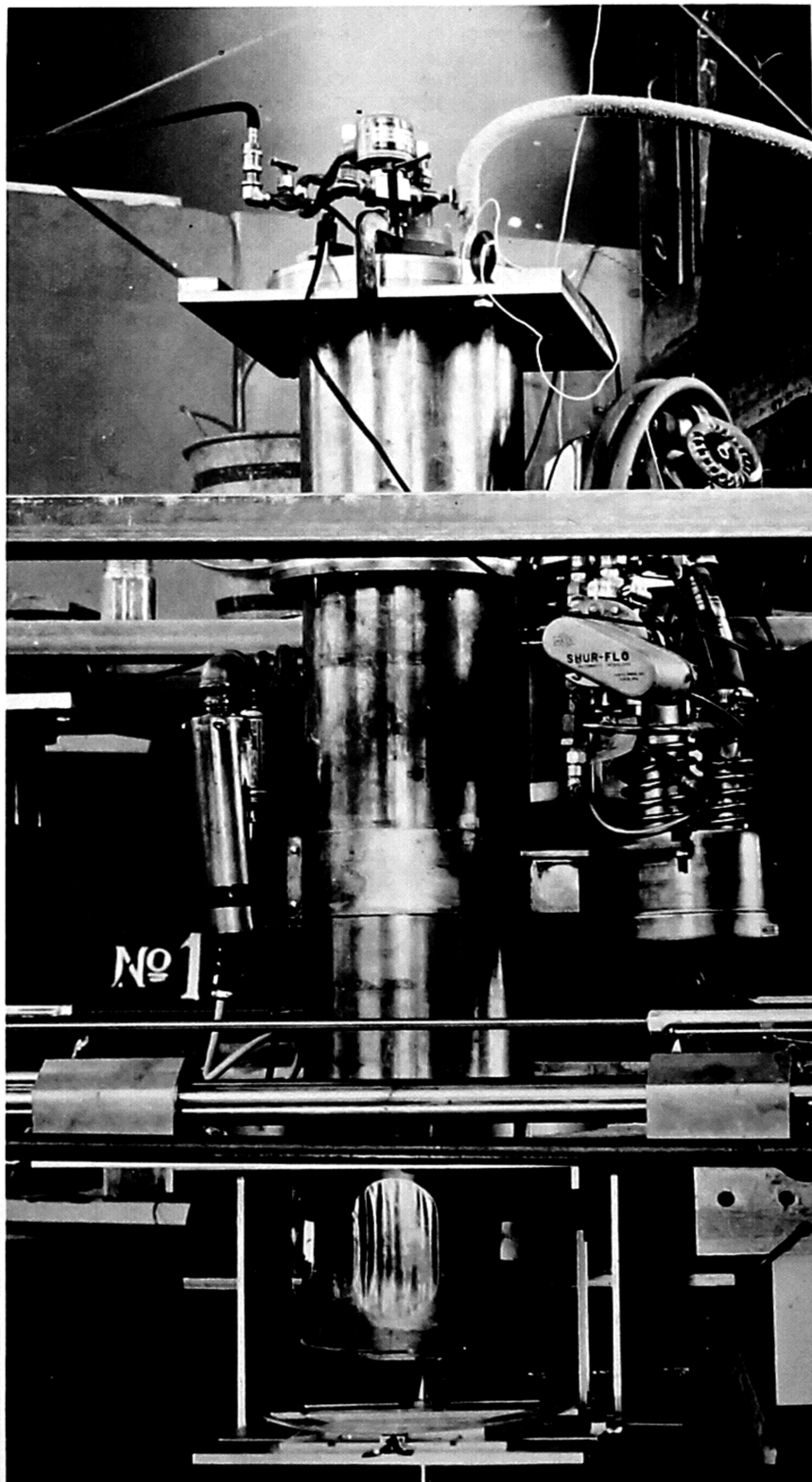
Pictures of the Force

Recently the method has been extended another step, and the asymmetrical beam has been scattered a third time. These triple-scattering experiments reveal still further details. They were first performed by the Nobel laureates Emilio Segrè and Owen Chamberlain and their colleagues at the University of California.

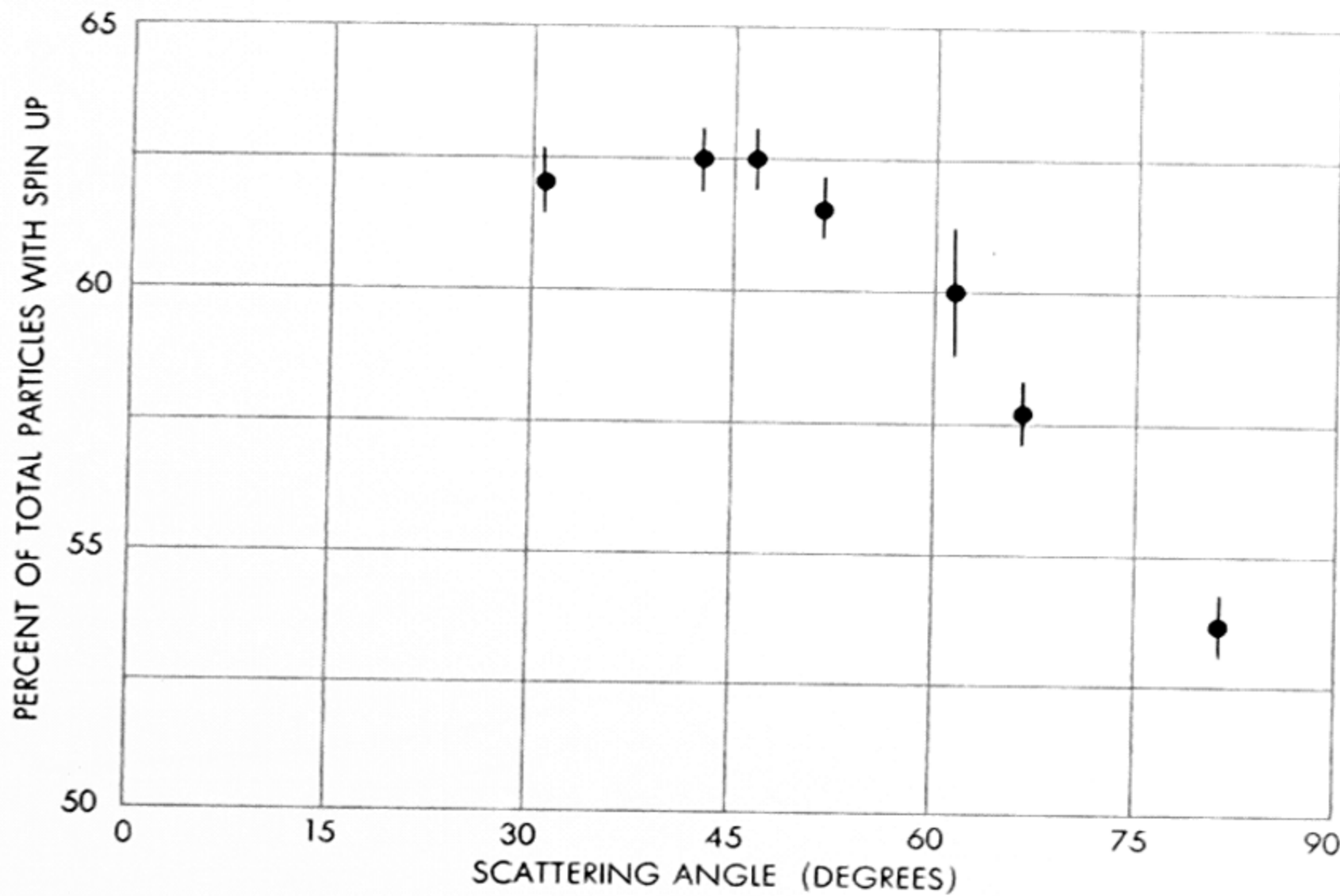
What are the results? From all that has been said it is obvious that we cannot speak of *the* nuclear force. We can only study all the nucleon combinations individually and determine the force for each. The one that has been most intensively studied so far is the proton-proton interaction. Having an electric charge, protons are easier projectiles to handle than neutrons. They can be accelerated, focused and aimed by electric and magnetic fields. Furthermore, there are targets of protons alone but not of neutrons alone.

At the bottom of the next page and on page 580 are samples of various components that combine to make up proton-proton forces. (The curves themselves show the potentials.) They represent a synthesis of calculations done by P. S. Signell and R. A. Bryan in collaboration with the author at the University of Rochester and by J. L. Gammel and R. M. Thaler at the Los Alamos Scientific Laboratory. We are fairly confident of the values of the potentials at distances greater than about one fermi, but are increasingly uncertain about the accuracy as the distance decreases. The bottom graph on page 579 shows the potential for the conditions of antiparallel spin, where the force has only a central component. The size of the force depends only on the distance between the particles. We see that at distances greater than .7 fermi the potential-energy curve slants up to the right, implying that the force is attractive. At shorter distances the curve slants up to the left, indicating a repulsive force. The force curve itself is a broken line.

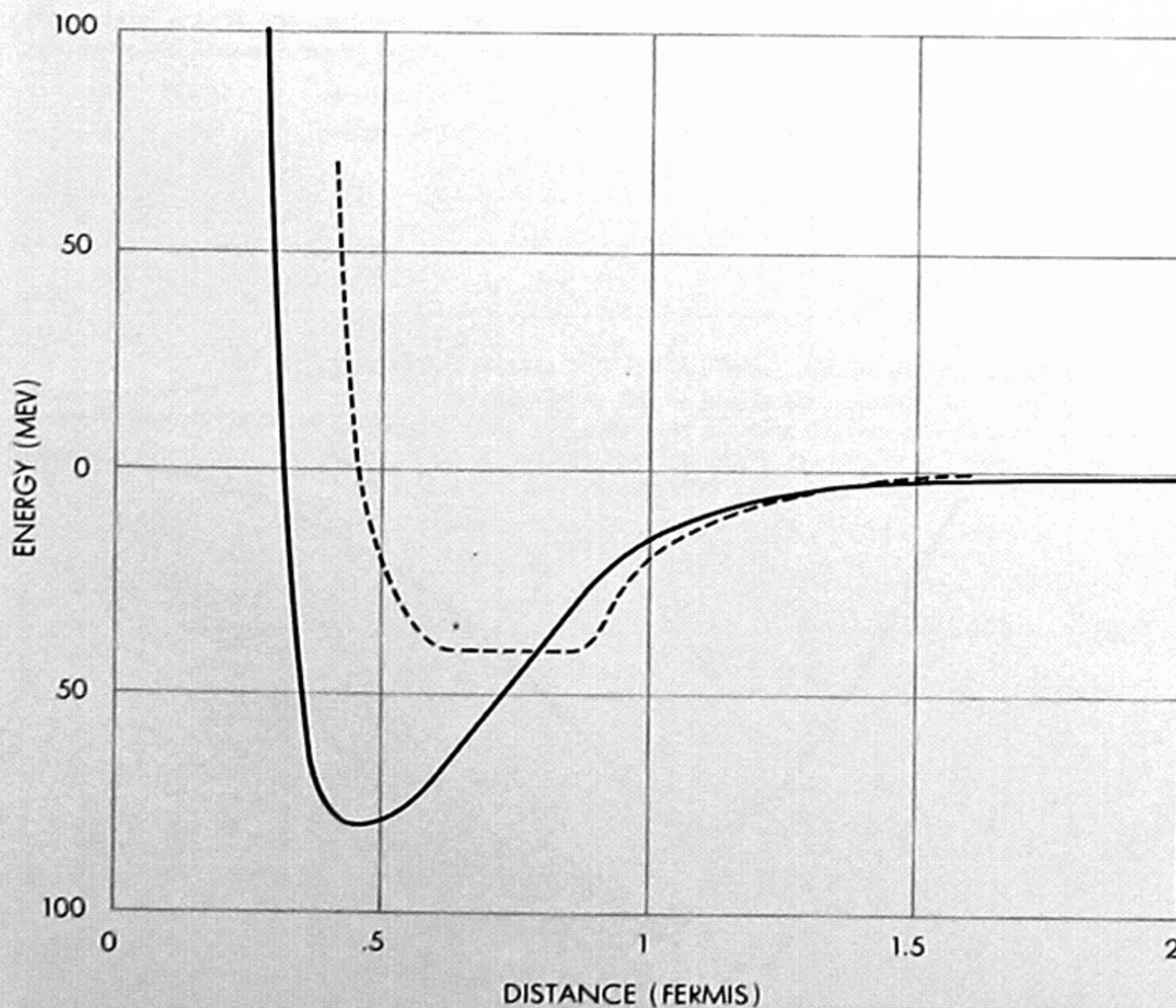
The top graph on page 580 contains



SECOND TARGET in triple-scattering experiment is liquid hydrogen contained in the brass cylinder photographed in close-up. Particles pass through the window at the bottom.



DEGREE OF POLARIZATION produced in scattering process varies with the scattering angle. Each point on this graph shows the per cent of total particles with spin vectors pointing up emerging at a specific angle from target. Lines represent limits of error.



PROTONS WITH ANTIPARALLEL SPINS can have only a central force. Corresponding potential energy is shown by solid curve. Broken curve indicates shape of force itself.

three curves, representing the potential energy due to the central force and one sample of the possible combinations of spin and orbital angular-momentum for the tensor and spin-orbit forces. We see that the central force differs for the antiparallel and parallel spin states. At larger distances the force for the antiparallel state is stronger (*i.e.*, the potential-energy curve slopes upward more sharply). The curve for the tensor potential represents the case where the particle-spins are at right angles to the orbital angular-momentum. The potentials are different when the spins are parallel and when they are antiparallel to the orbital angular-momentum. The same is true for the spin-orbit curve. Both types of force are strong and of short range, but the range of the spin-orbit force is about half that of the tensor force. This means that higher energy is required to bring the spin-orbit force into play than is required for the tensor force.

So we find that the nuclear force between two protons is about as complicated as it could possibly be. Two types of central force, depending on the relative orientation of the spins, and two types of noncentral force are necessary to explain all the experimental facts. And the noncentral forces even change their signs and magnitude depending on the relative orientation of the spin and orbital angular-momenta.

What about neutron-neutron forces? Since there are no pure neutron targets, the experiments must necessarily be less direct. Neutrons are scattered from targets such as deuterons, containing both protons and neutrons, and the effect of the pure neutron-neutron force must be calculated. So far as we can tell, these experiments confirm the long-suspected principle of charge symmetry. Nature has made an astonishing choice in favor of simplicity. With a clear opportunity to distinguish between protons and neutrons, she has chosen not to do so. The nuclear force between two neutrons appears to be exactly the same as that between two protons.

Furthermore, the independence of the nuclear force with respect to charge seems to extend to the neutron-proton system, but with a new and complicating feature added. Because the neutron and proton are different particles, the Pauli principle does not apply to them. Therefore they can go together in combinations that are excluded for two protons and two neutrons. Specifically, they can have antiparallel spins together with odd orbital angular-momentum, and parallel

spins together with even orbital angular-momentum.

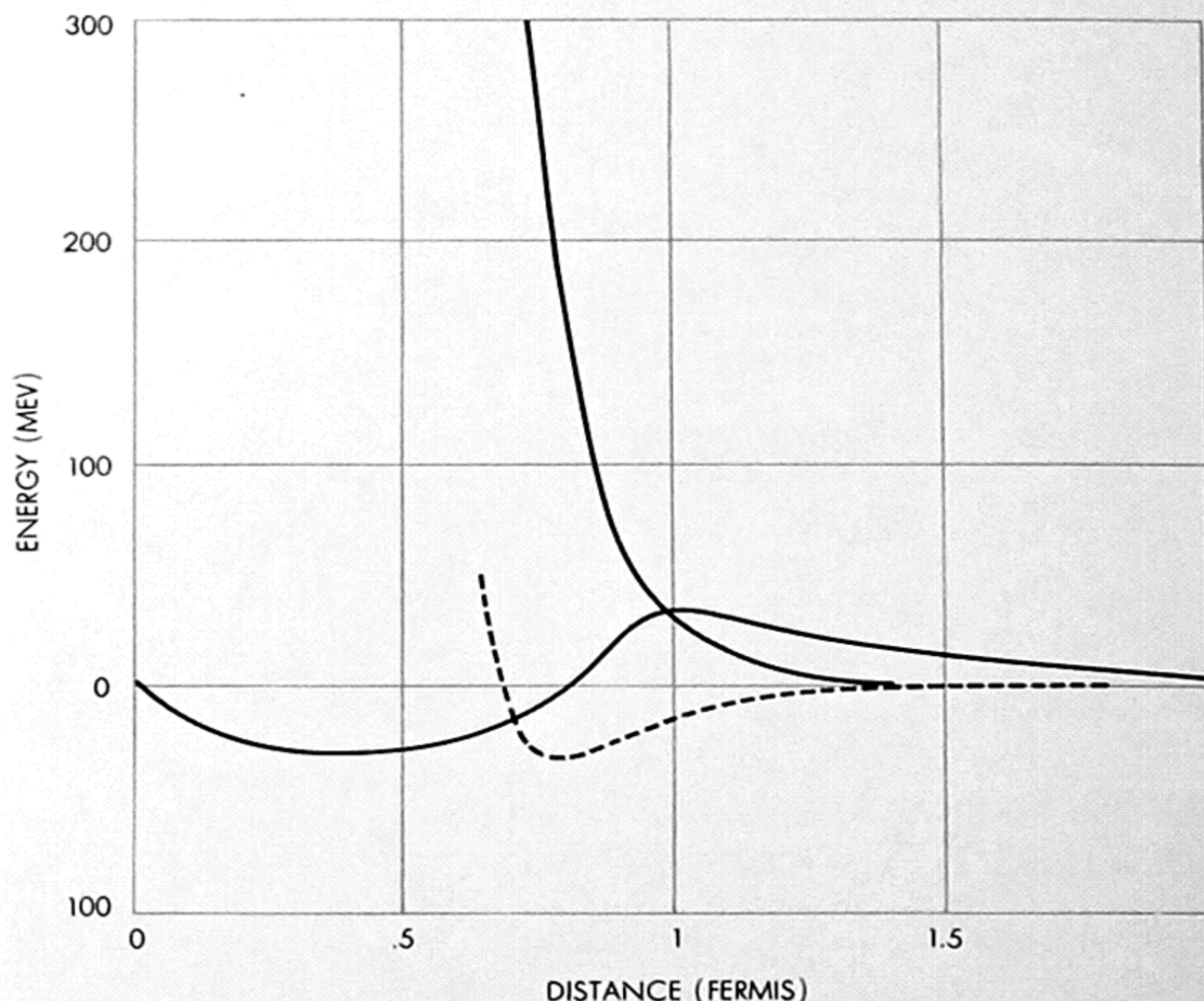
Unfortunately no triple-scattering experiments for neutron and proton have been carried out so far; they are very difficult. But there have been single- and double-scattering experiments for quite a range of energies. The lack of triple-scattering data and the fact that there are more possible states have made it much more difficult to pin down the nuclear force between the neutron and the proton.

Analyses of the available scattering results, and of the properties of the deuteron (which is a bound neutron-proton system with parallel spins and even orbital angular-momentum) suggest that the neutron-proton force is also the same as that between two protons for the states they share in common (even orbital angular-momentum and antiparallel spins; odd orbital angular-momentum and parallel spins). The potential energies for the other two combinations of states are different. The bottom graph on this page summarizes what we know about them. Note that a tensor force is the only noncentral type represented. We do not yet have enough information to decide whether there is a spin-orbit force as well.

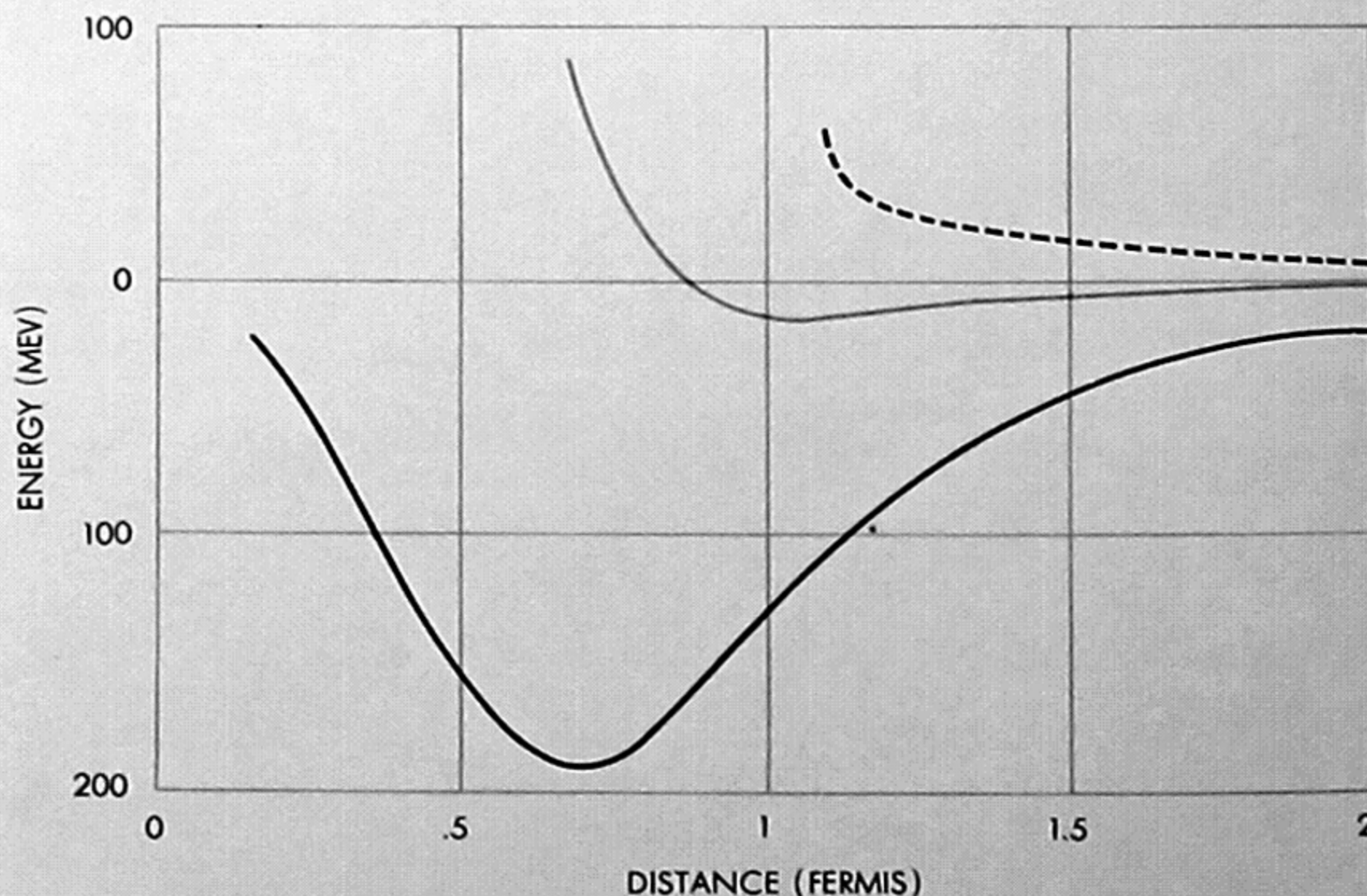
Working with these recently discovered potential energies between two nucleons, theorists are having increasing success in explaining the properties of nuclei larger than the deuteron. As the force picture is further sharpened, nuclear physics will continue to advance.

What of the dream of a true theoretical understanding of the nuclear force? At the moment we can hardly do more than hope we are on the right track. There seems every reason to suppose that Hideki Yukawa's famous conjecture was right, and that the pion is the "agent" that is responsible for the force between two nucleons. But although we can write equations based on this idea, we do not know how to solve them. Therefore we cannot even be sure that the equations themselves are right. Very rough approximate solutions do seem to give the correct nuclear force for large distances. For smaller distances, however, the answers do not agree with the experimental results.

We are still confident that some day we shall be able to write down the correct equations and find a way of solving them. Meantime there seems no choice but to push ahead with the empirical approach described in this article.



PROTONS WITH PARALLEL SPINS have a central force (*broken curve*), a tensor force (*black curve*) and a spin-orbit force (*gray curve*). The curves actually show potential energy rather than force itself. The tensor and spin-orbit curves that are drawn here apply to the case where the spin vector is perpendicular to the orbital angular-momentum vector.



FORCE BETWEEN PROTON AND NEUTRON can have components that are ruled out in proton-proton and neutron-neutron cases by the Pauli exclusion principle. Dashed curve gives potential energy for antiparallel spins and odd orbital angular-momentum (*l*). Gray curve applies to the central force for parallel spins and even *l*; black curve, to the tensor force for the same situation. The existence of a spin-orbit force is still uncertain.

The Author

ROBERT E. MARSHAK is Harris Professor of Physics and chairman of the physics department at the University of Rochester. He was born in New York City in 1916 and attended Columbia University as a Pulitzer scholar. He graduated at the age of 19 and three years later took a Ph.D. in theoretical physics under Hans A. Bethe at Cornell University. Later he was associated with Victor F. Weisskopf at Rochester, and during the war he worked for the Manhattan District. He has written three other articles for SCIENTIFIC AMERICAN.

Bibliography

QUANTITATIVE THEORY OF NUCLEAR FORCES. H. A. Bethe in *Elementary Nuclear Theory*, pages 23-96. John Wiley & Sons, Inc., 1947.

TWO-BODY PROBLEMS AT HIGH ENERGIES. John M. Blatt and Victor F. Weisskopf in *Theoretical Nuclear Physics*, pages 168-190. John Wiley & Sons, Inc., 1952.

TWO-BODY PROBLEMS AT LOW ENERGIES. John M. Blatt and Victor F. Weisskopf in *Theoretical Nuclear Physics*, pages 48-118. John Wiley & Sons, Inc., 1952.

THE TWO-NUCLEON PROBLEM. Lamek Hulthen and Masao Sugawara in *Encyclopedia of Physics*, Vol. 39, pages 1-143; 1957.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

APPLICATIONS OF SUPERCONDUCTIVITY

by Theodore A. Buchhold

When certain metals are cooled almost to absolute zero, they become superconductive, that is, their electrical resistance disappears. This remarkable phenomenon is now being turned to technological purposes.

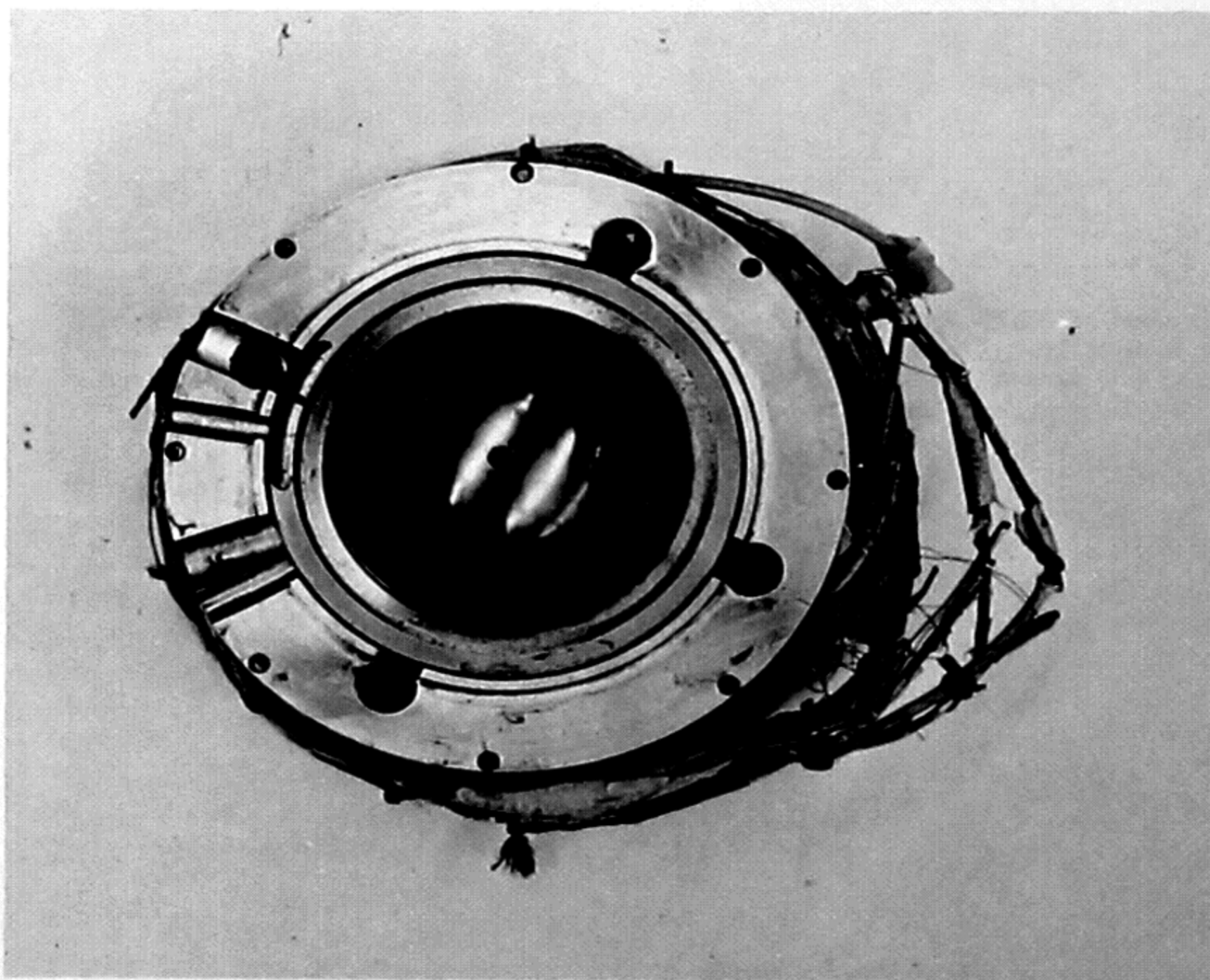
Superconductivity—the total disappearance of electrical resistance—is a strange property of matter that makes its appearance in the remote realm of temperatures close to absolute zero. It confounded the classical theory of electromagnetism when it was discovered some 50 years ago. It remained until four years ago an intractable riddle to the new quantum physics. Yet in a very few years from now superconductivity is going to prove to be a windfall to technology. Engineers are engaged in the

design of superconductive bearings distinguished by the absence of friction, electric motors with extraordinary efficiency, tiny and reliable switching elements for computer circuits, magnetic lenses with unprecedented resolving power for electron microscopes, noiseless amplifiers and other devices with characteristics that approach ideal standards. In the future workers in other fields of engineering are sure to find that superconductors, insulated from their immediate environment in cold

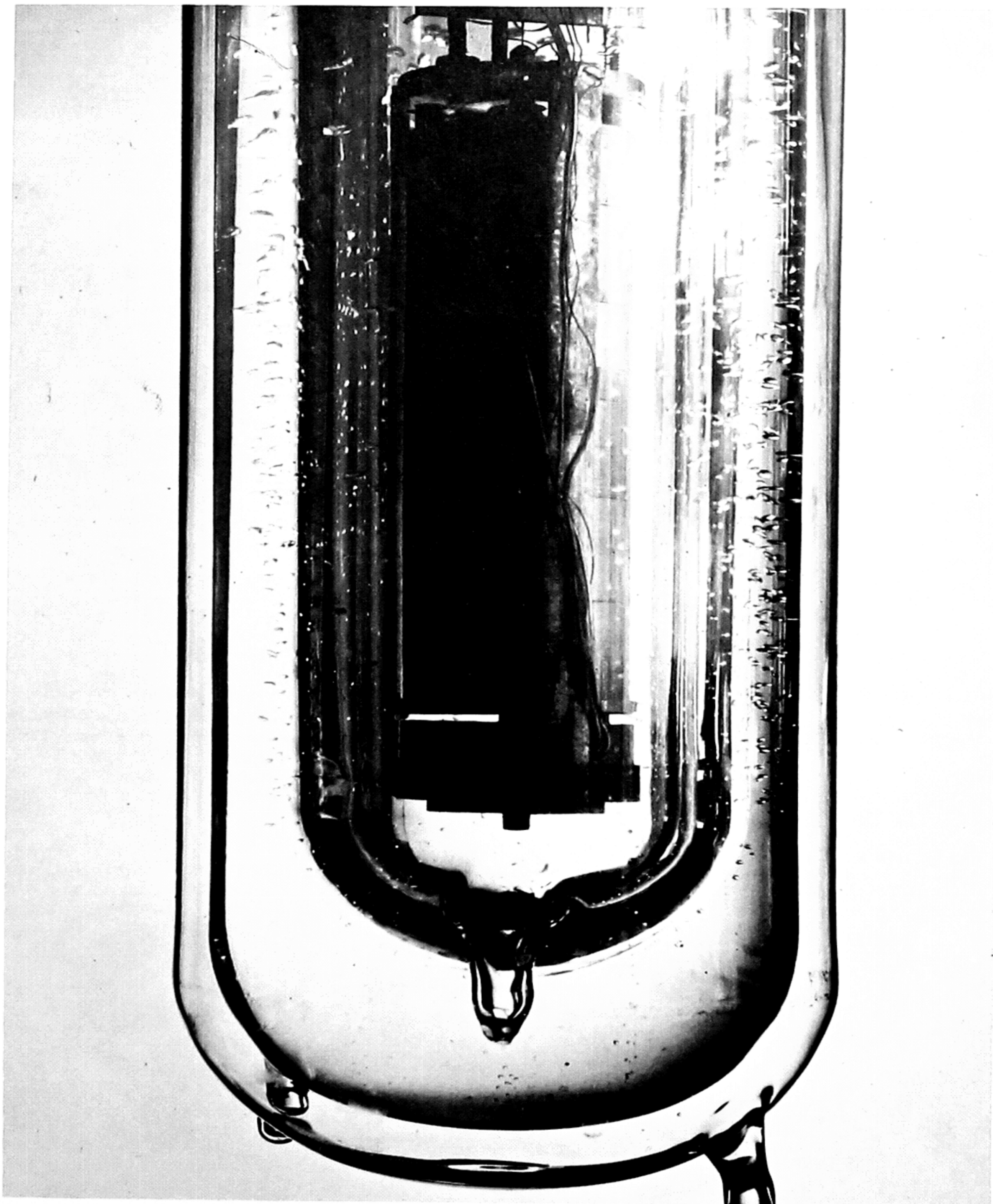
chambers, will provide unexpected solutions for an increasing variety of design problems.

It was well understood in classical electromagnetism that the resistance of an electrical conductor should fall with temperature. The explanation could be pictured quite graphically. An electric current consists in the flow of free electrons through the crystal lattice of the conductor. At room temperature the thermal vibration of the atoms in the lattice increases the probability that the electrons will collide with them; this impedes the flow of electrons and sets up resistance to the current. At lower temperature, with the vibration of the atoms reduced in amplitude, the electrons collide with the atoms less frequently, and the current encounters less resistance. At absolute zero the atoms were supposed to cease vibrating entirely. But there would remain some resistance to the flow of current because a few electrons would still collide with the now-stationary lattice and with the defects and impurities that distort the lattice structure in all but perfect crystals.

This model worked with complete satisfaction down the temperature scale as investigators approached closer and closer to absolute zero. Then, in 1911, it failed completely when the Dutch physicist Heike Kamerlingh Onnes froze mercury in a bath of liquid helium. He ran a current through the mercury and watched as his instruments showed resistance declining with temperature. The familiar relationship between the two properties held until the temperature reached 4.2 degrees absolute (degrees centigrade above absolute zero). Suddenly the electrical resistance of the mercury vanished; there was not even the residual resistance betokening collisions between electrons and defects and

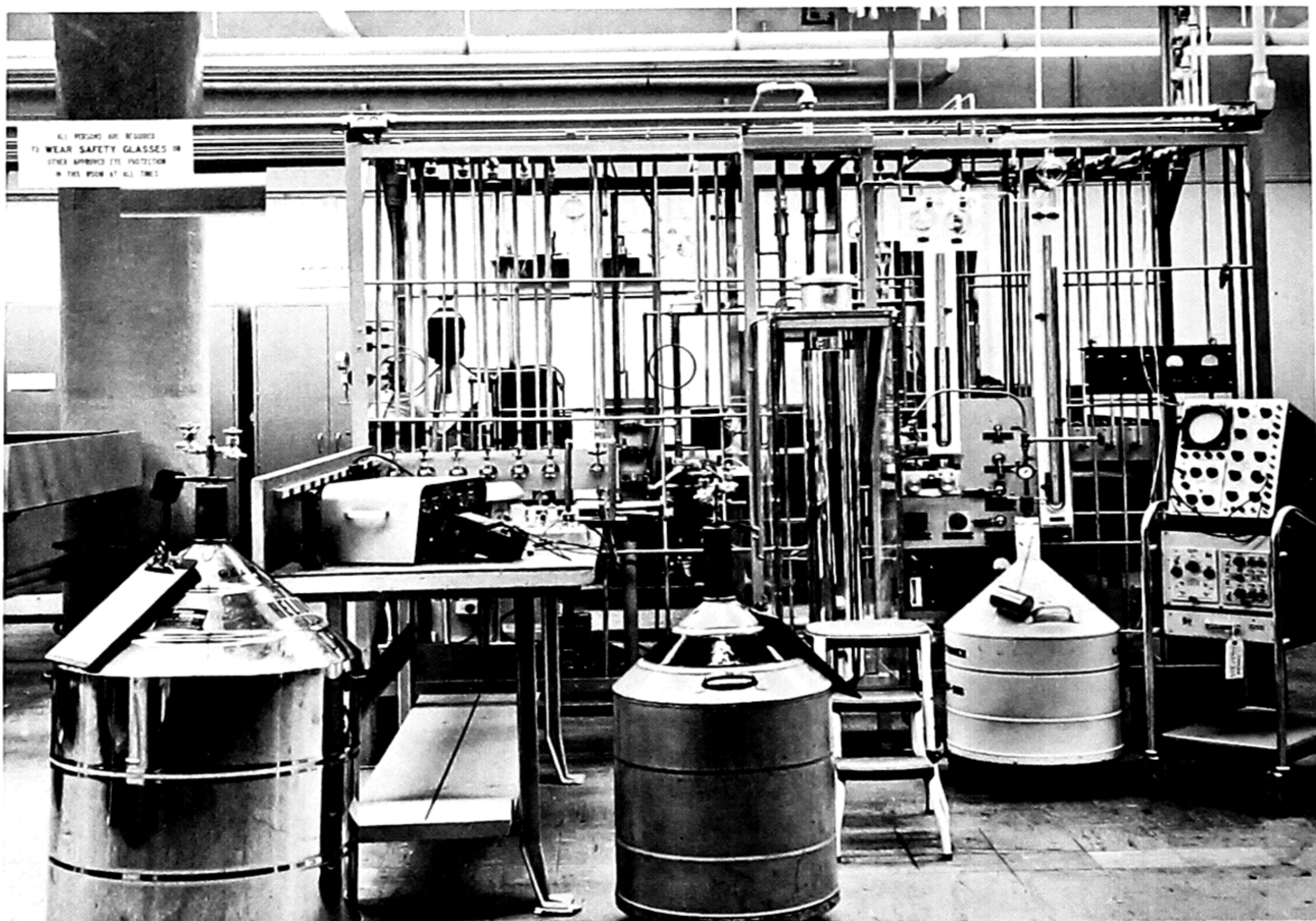


SUPERCONDUCTIVE GYROSCOPE, shown here in pilot model, is being developed by the General Electric Company. The spherical rotor floats without friction in an "atmosphere" of magnetic flux. Upper housing has been removed to show interior structure.



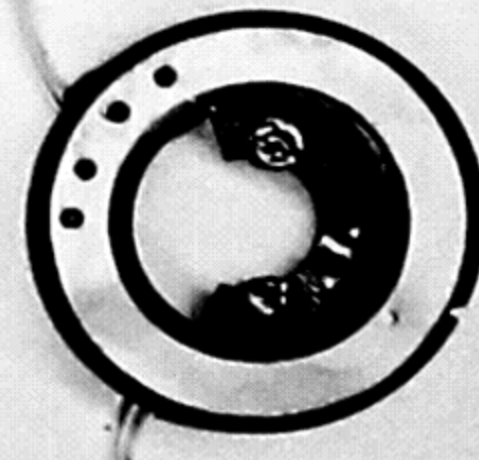
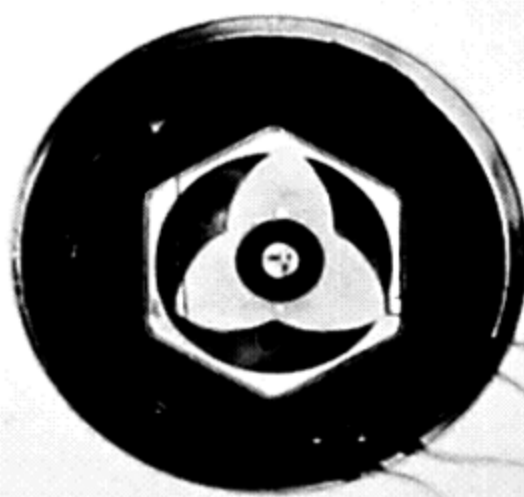
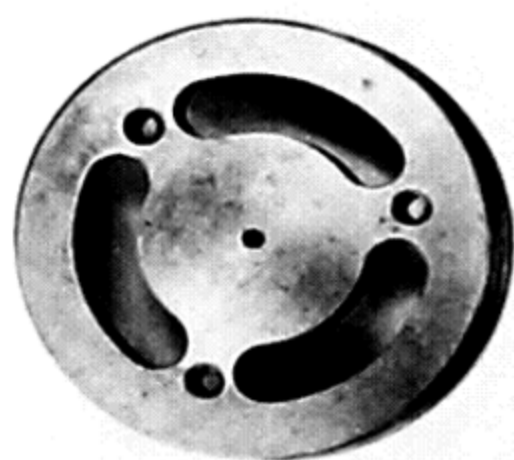
LOW-TEMPERATURE ENVIRONMENT is maintained in this Dewar flask within a Dewar flask. Outer double-walled flask holds liquid nitrogen (at 77.3 degrees absolute), which insulates liquid

helium (at 4.2 degrees) in the inner double-walled flask. Tall cylinder, weighing 12 pounds, is supported by magnetic flux of small cylinder at bottom. Bubbles are due to boiling of liquids.



LOW-TEMPERATURE INSTALLATION at the General Engineering Laboratory of the General Electric Company is the site of

the experimental apparatus depicted on the preceding page. The large Dewar flask is at right center. Cans contain liquefied gases.



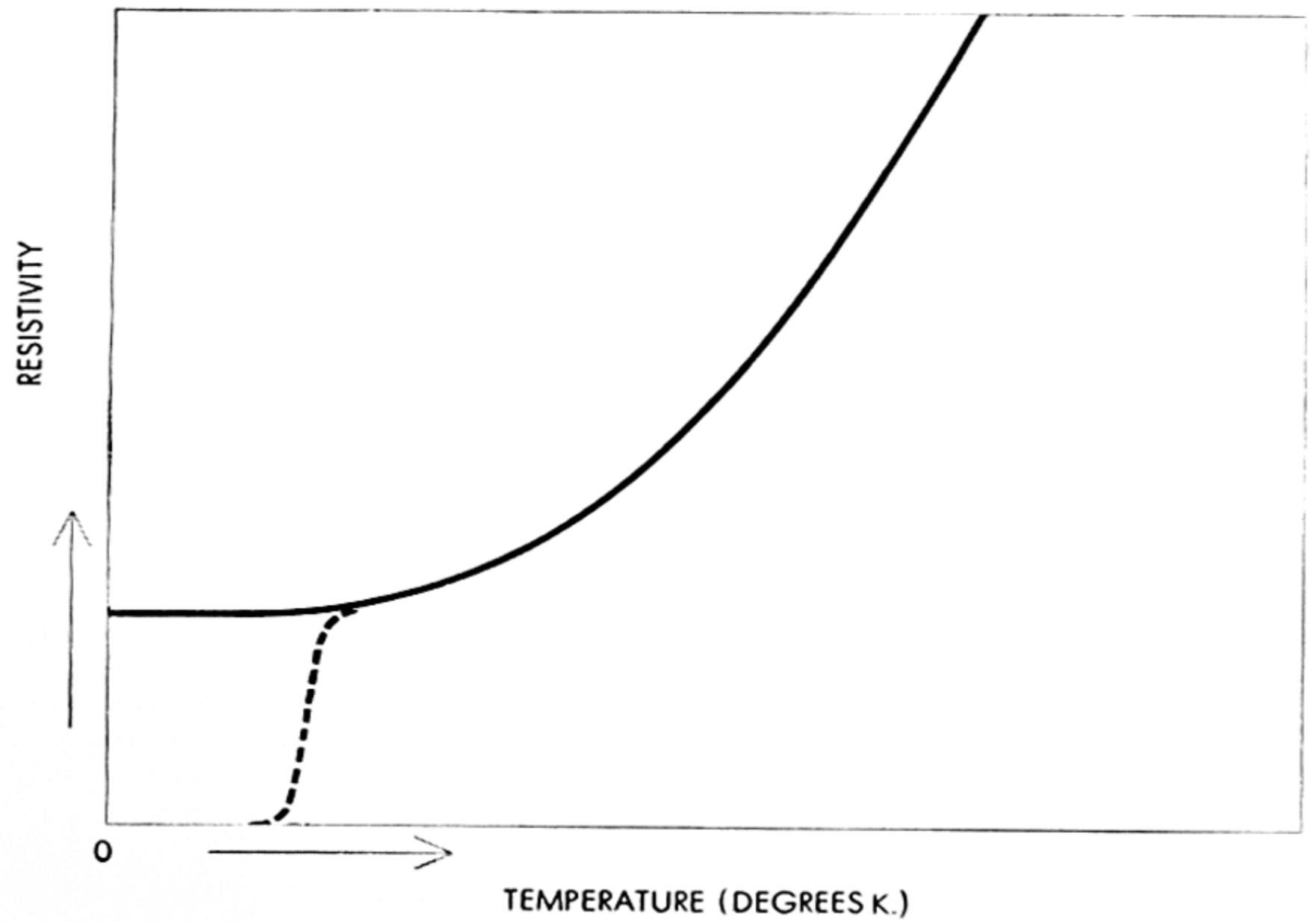
PROTOTYPE OF SUPERCONDUCTIVE MOTOR includes supporting elements (*left and right*) and stator housing (*middle*). Hexagonal structure in stator housing is the rotor, an aluminum

core wrapped in niobium foil, which turns at 700 revolutions per minute in a bath of liquid helium. Its hexagonal shape, necessary for it to develop torque, is explained in illustration on page 589.

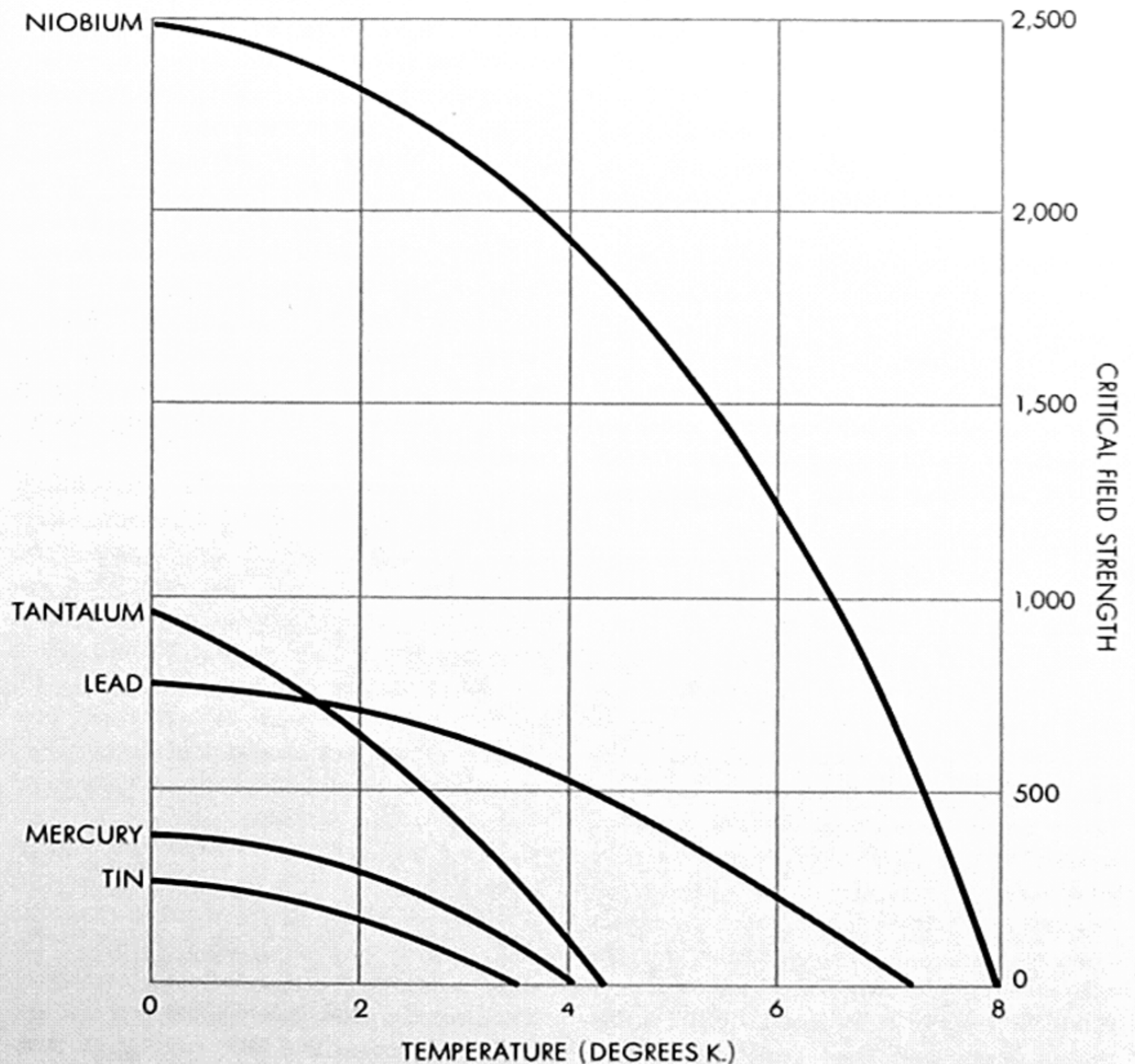
impurities in the lattice predicted by the classical model.

Kamerlingh Onnes found that other metals, among them tin, lead, tantalum and niobium display the same abrupt disappearance of resistance near absolute zero. It does not occur in all metals; strangely, the best electrical conductors, copper and silver, do not become superconductive. But certain alloys and compounds, as well as pure metals, become excellent superconductors under the right conditions. The property puts in its appearance and disappearance just as suddenly in each material but at a different "transition" temperature—generally only a few degrees above absolute zero. Superconductivity can also be destroyed if the metal or alloy is exposed to a magnetic field of sufficient strength, whether the field is applied externally or set up by the current flowing through the superconductor itself. In each material the field strength required to erase superconductivity varies with temperature within the range in which the material is superconductive. The metal niobium, for example, has a transition temperature of 8 degrees; its critical field-strength is 2,000 oersteds at 4.2 degrees and rises to 2,400 oersteds at 1 degree. Critical field-strength varies not only with temperature, but also with the purity of the material, with mechanical stress and with the configuration of the specimen in question. Depending upon these factors, niobium has shown field-strengths as high as 4,000 oersteds. An alloy of niobium and tin has so far shown the highest transition temperature, 18 degrees, and may have a critical field-strength as high as 16,000 oersteds at 4.2 degrees.

Experimental findings of this kind troubled theoretical physicists for more than 40 years. Superconductivity was the one phenomenon in the domain of quantum physics which that powerful system could not explain. Some six years ago B. T. Matthias of Bell Telephone Laboratories set out to discover by empirical methods the order underlying the appearance of superconductivity in elements and compounds. He made systematic measurements on a great variety of materials, and in the course of his investigation developed many promising new superconductors. He also found a pattern of results that sometimes helps in predicting which materials will become superconductive. In general it appears that elements with three, five or seven electrons in their outer electron-shells become superconductive most easily,



ABRUPT DISAPPEARANCE OF RESISTANCE in superconductors as they approach absolute zero (-273 degrees centigrade) is indicated by the broken line in this graph. The resistance of nonsuperconductive metals falls to a minimum value (solid curve).



CRITICAL FIELD STRENGTH of superconductors rises as temperature falls. Niobium (top curve) becomes superconductive at about 8 degrees absolute, but at that temperature even the weakest magnetic field renders it resistive again. At 5 degrees niobium withstands a field strength of some 1,600 oersteds; at 1 degree, a field strength of almost 2,500 oersteds.

while those with only one or with more than eight do not become superconductive at all [see "Superconductivity," by B. T. Matthias; SCIENTIFIC AMERICAN Offprint 227].

As Matthias proceeded on this line of investigation, John Bardeen and his colleagues at the University of Illinois succeeded in bringing superconductivity within the equations of quantum theory. They found that some fraction of the total population of current-carrying electrons in a superconductor are "paired" in the sense that the resistance set up by the collision of one electron is precisely offset by the rebound of its partner from a simultaneous collision, so that no net resistance to the current is set up. At temperatures above the transition point, or in magnetic fields of greater than critical strength, these electrons are "unpaired." Their collisions are no longer self-canceling but additive, and electrical resistance is restored.

The first practical applications of superconductors arise not so much from the currents they conduct so easily as from the magnetic fields these currents set up. Once a current has been started in a superconductor it continues to flow without variation or decay—theoretically forever—after the power source has been disconnected [see top illustration on page 588]. The superconductor becomes a kind of perpetual-motion machine. But it is the magnetic flux, frozen into the superconductor with the current, that makes this curious phenomenon useful. The frozen-in flux makes it possible to use the current to calibrate other currents. The current to be calibrated is simply conducted through the cold environment, and its magnetic flux is compared with that of the superconductor.

The same relationship between electricity and magnetism accounts for another remarkable property of superconductors: they act as insulators against magnetic fields. When an external magnetic field is brought near a superconductor, or any conductor for that matter, it sets up eddy currents in a surface layer of the material. In an ordinary conductor this surface layer may be relatively deep, and the resistance of the material quickly suppresses the eddy currents. In a superconductor, on the other hand, the eddy currents are confined to a layer only about .0001 millimeter deep, and they continue to flow without deterioration. The magnetic flux associated with the eddy currents is oriented in a sign parallel to that of the external field. In effect the surface of

the superconductor becomes a magnetic mirror, reflecting the lines of force of the impinging field.

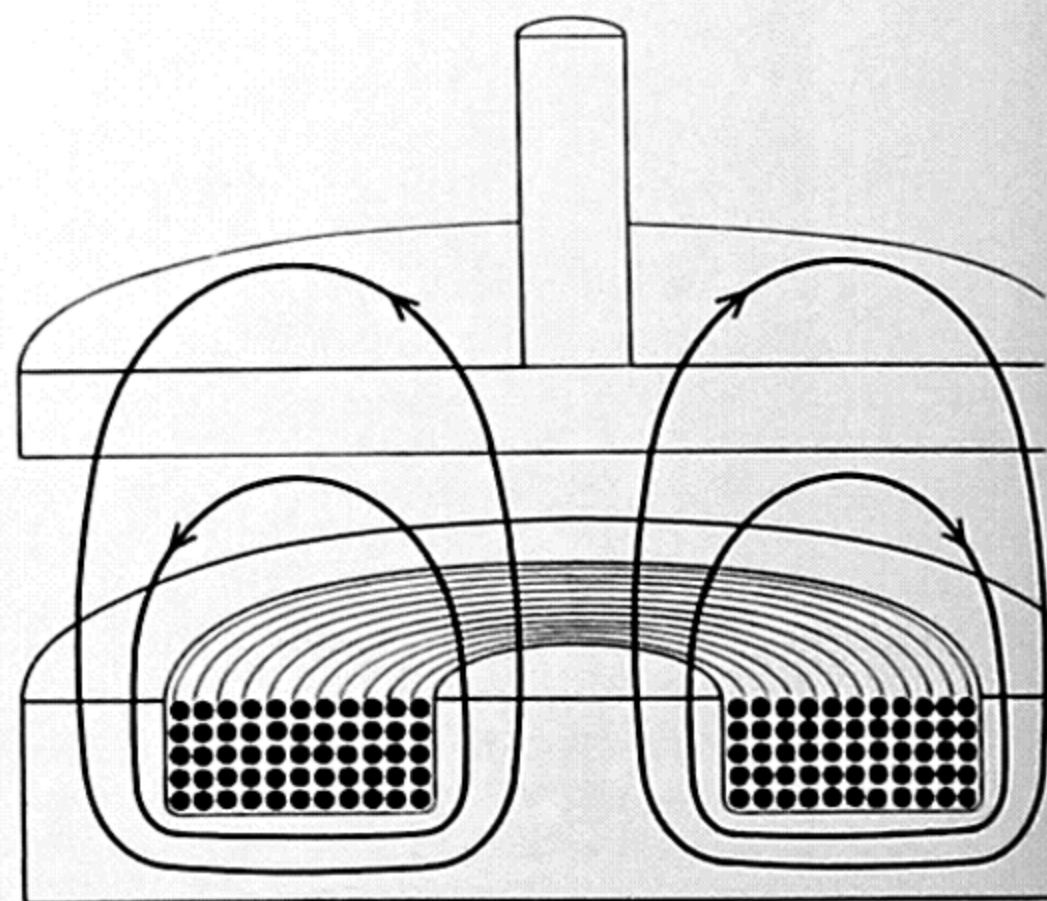
The magnetic-mirror effect suggests a variety of useful applications of superconductivity. In an electron microscope, for example, magnetic fields focus the image formed by the beam of electrons. These fields are shaped to act as "lenses" by the doughnut shape of the magnet that produces them, the lines of force being concentrated in the hole in the middle. Since the magnetic-mirror effect involves only the outer surface of materials, a superconducting foil shielding the contours of the center hole and employed as shielding elsewhere in the instrument will perfect the shape of the magnetic lens. The improvement in resolution may make it possible to produce an electron microscope in which atoms can be seen. For certain biological and chemical investigations (for example, studies of free chemical radicals) which require low temperatures in any case, the necessary refrigeration will create no serious complications.

Superconductive surfaces might also improve the performance of the resonant cavities of the oscillators that generate microwaves. Because superconductors offer little resistance to high-frequency electric currents, they would greatly enhance the efficiency of these cavities and stabilize the oscillator on its frequency with an accuracy approaching that of an atomic clock. In the low-temperature environment necessary for superconductivity, moreover, metals neither expand nor contract; the enhanced structural stability would hold the geometry of the resonating cavity constant.

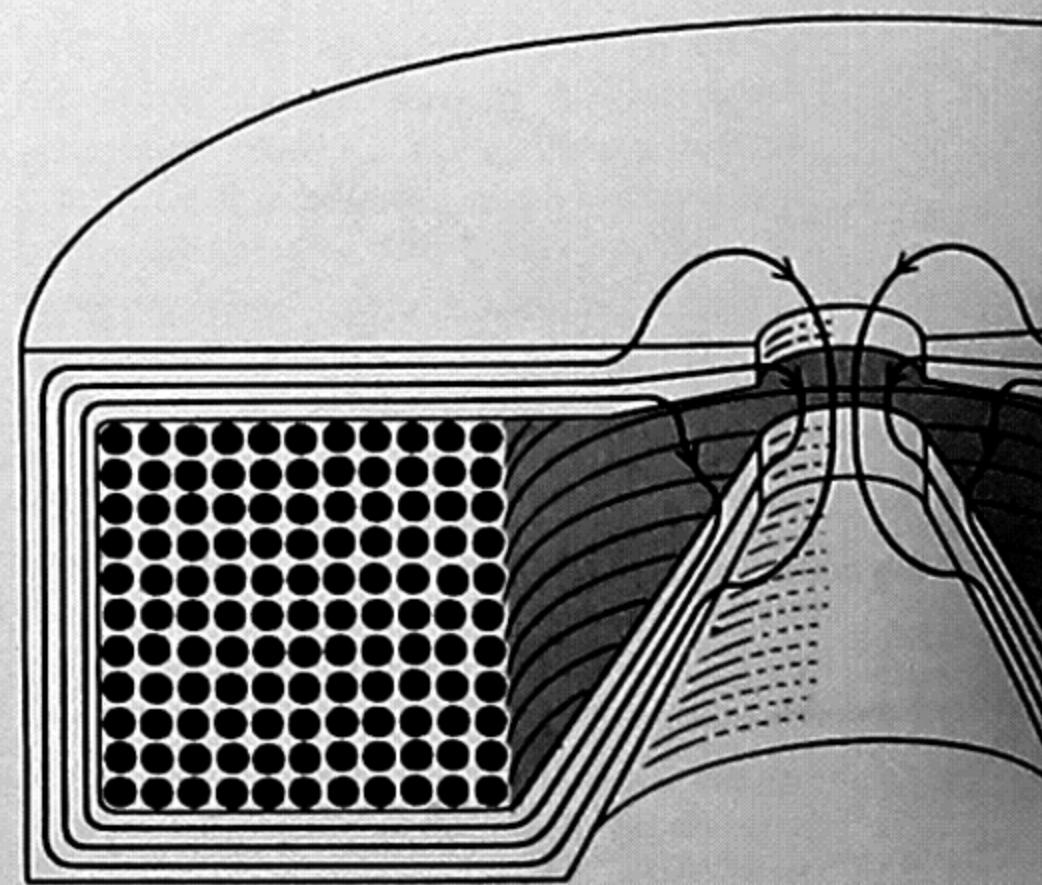
The magnetic-mirror effect inevitably suggests the idea of a frictionless magnetic bearing, one that would ride upon a cushion of magnetic flux. At the General Engineering Laboratory of the General Electric Company we have made a model of such a bearing. It consists of a superconductive coil of niobium wire above which we mount a superconductive disk in such a way that it can move vertically but not laterally. A current passing through the coil creates a magnetic flux that repels the disk and pushes it upward. With the flux trapped in the coil [see top illustration at right], the disk remains suspended on the flux. Pushing the disk downward toward the coil compresses the flux, raising its density, thereby increasing the current in the coil and amplifying the repelling force of the flux; the upward force of the flux increases, in fact, as the square

of the increase in the flux density. Tests have shown that the model bearing can produce an upward push of 300 grams per square centimeter of coil surface. With small gaps between disk and coil, good bearing stiffness can be obtained. Conventional magnetic and electrostatic devices cannot match this performance without the aid of complicated feedback circuitry.

It is only a small leap of the imagination from a simple bearing such as this, which sustains a vertical load, to a



MAGNETIC BEARING is based upon the ability of superconductors to insulate magnetic fluxes. In the illustration at left the magnetic flux (black lines) generated by



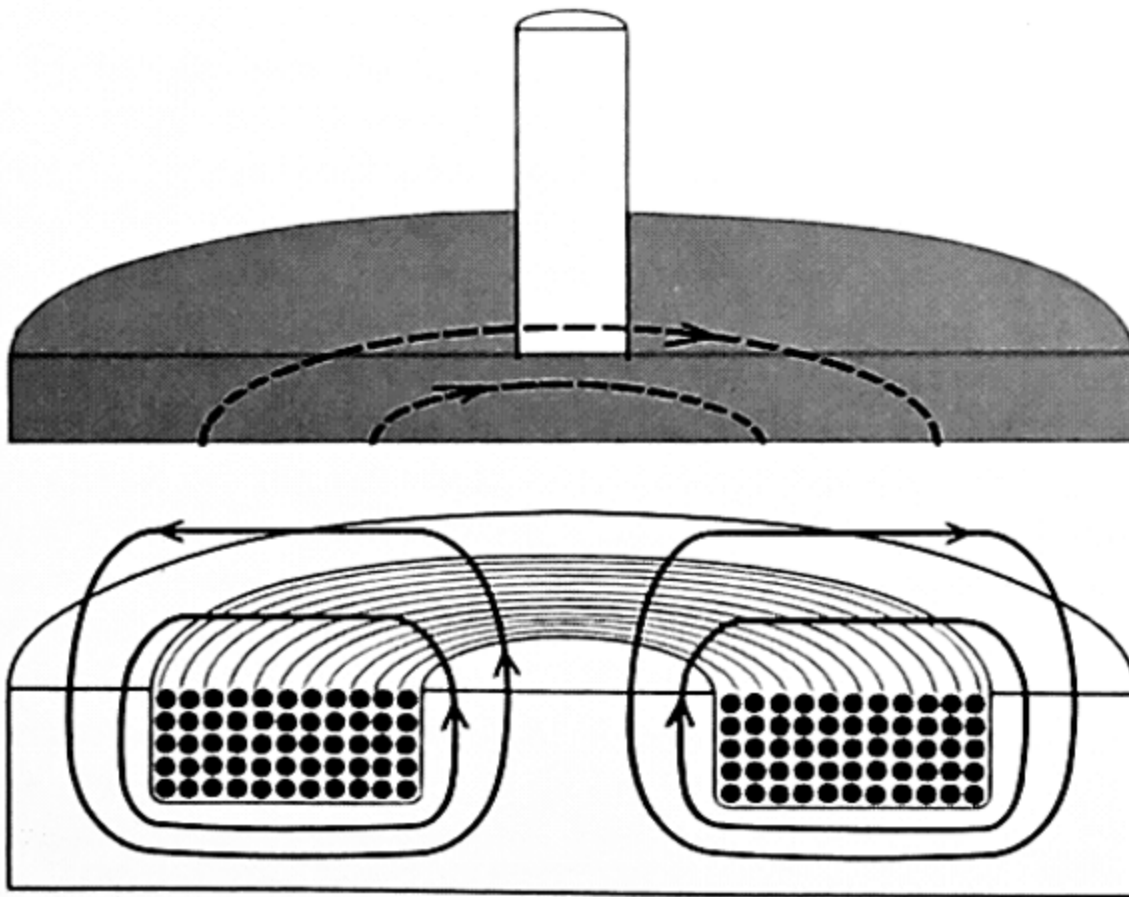
MAGNETIC-INSULATING SHIELDS may so greatly enhance the resolving power of electron-microscope lenses that they could

pair of magnetic bearings to support a rotating shaft [see *bottom illustration on page 588*]. In another leap one can conceive of a three-dimensional arrangement of six such bearings to float a body in a magnetic "atmosphere." If the floating body is now rotated, it becomes in effect a gyroscope in which the superconductive magnetic bearings replace the conventional gimbals. To make it spin, the body would of course become the rotor of an electric motor.

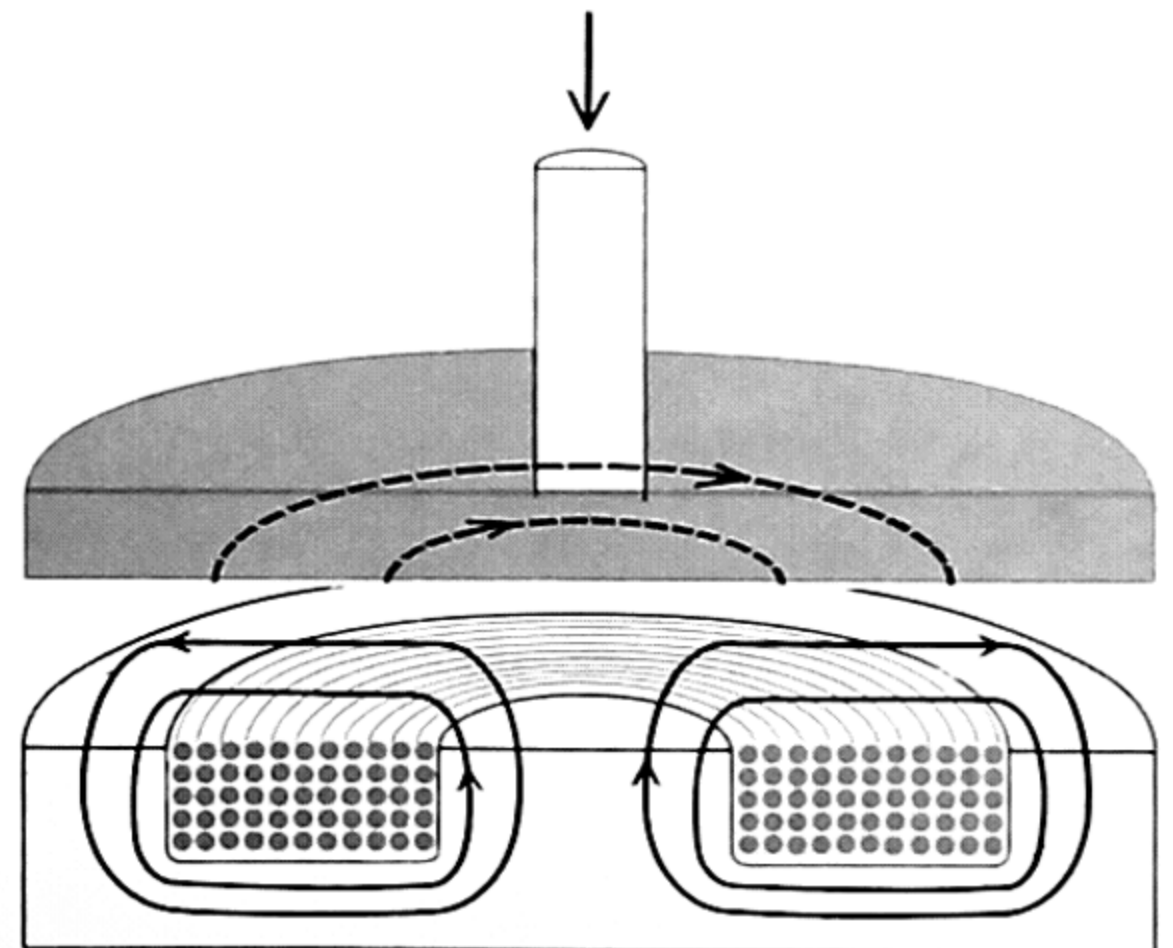
The superconductive electric motor is a topic in itself. Superconductive bear-

ings would eliminate the major source of friction. If the stator coils as well as the rotor are made superconductive, the motor should have an efficiency of nearly 100 per cent. The rotor, however, cannot have the conventional cylindrical shape. Since the lines of force set up by the stator coils are perpendicular to the surface of the rotor, they produce no torque, and the rotor will not turn on its bearings. If the rotor is given the shape of a polygon, the lines of force exert the desired torque [see *top illustration on page 589*]. A rotating flux,

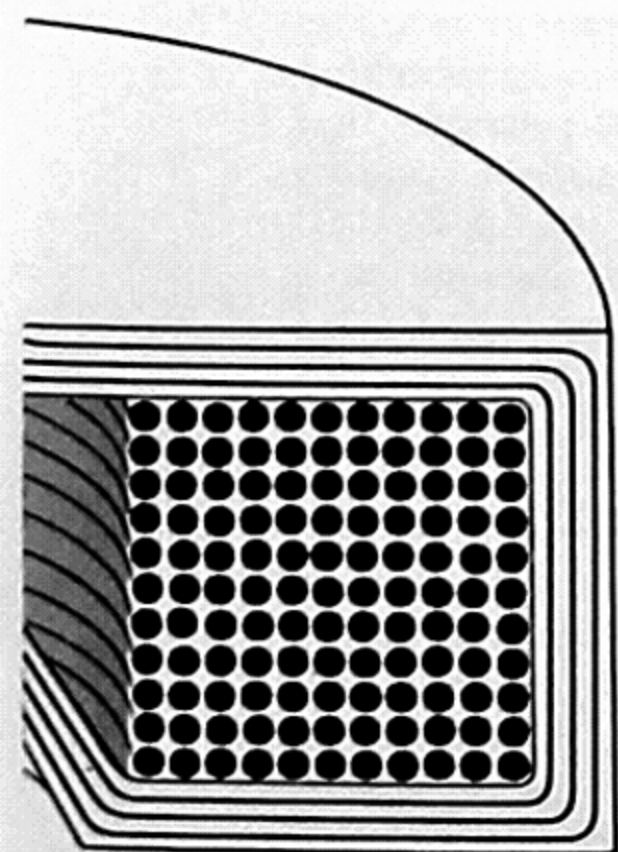
produced as in any alternating-current polyphase winding motor, will turn the rotor; increase in the frequency of the coil currents will turn the rotor faster. We have built a model motor on these principles and driven it up to 20,000 revolutions per minute, the speed being limited only because the rotor was not built to withstand higher centrifugal forces. The low temperature in which the motor must operate to maintain superconductivity limits its usefulness to special applications; for example, a high-precision gyroscope.



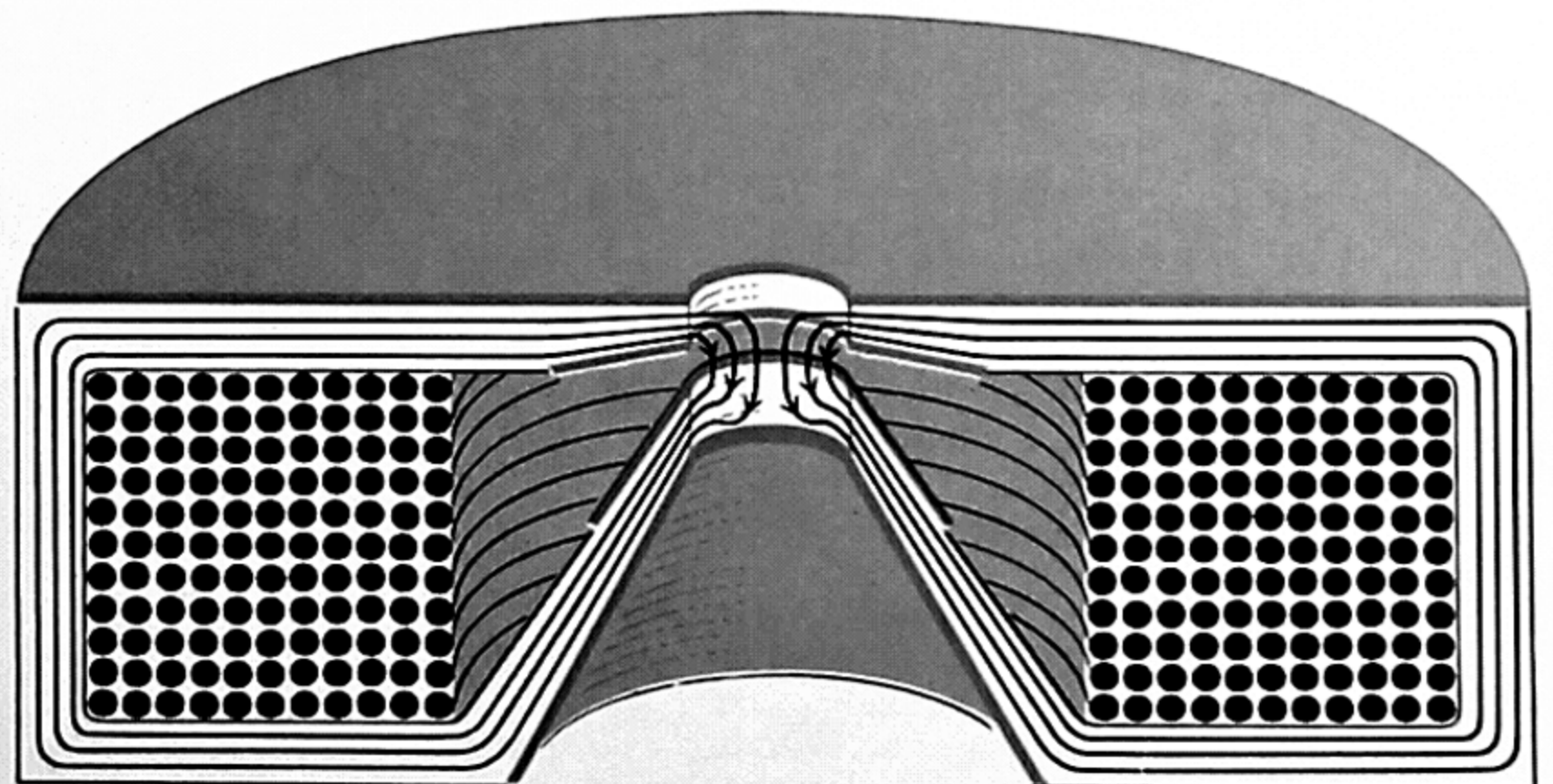
the electromagnetic coil passes unhindered through a disk which is not superconductive. The magnetic insulating ability of a superconductive disk (*middle illustration*) arises from its ability to repulse the magnetic flux. In the illustration at right this ability to



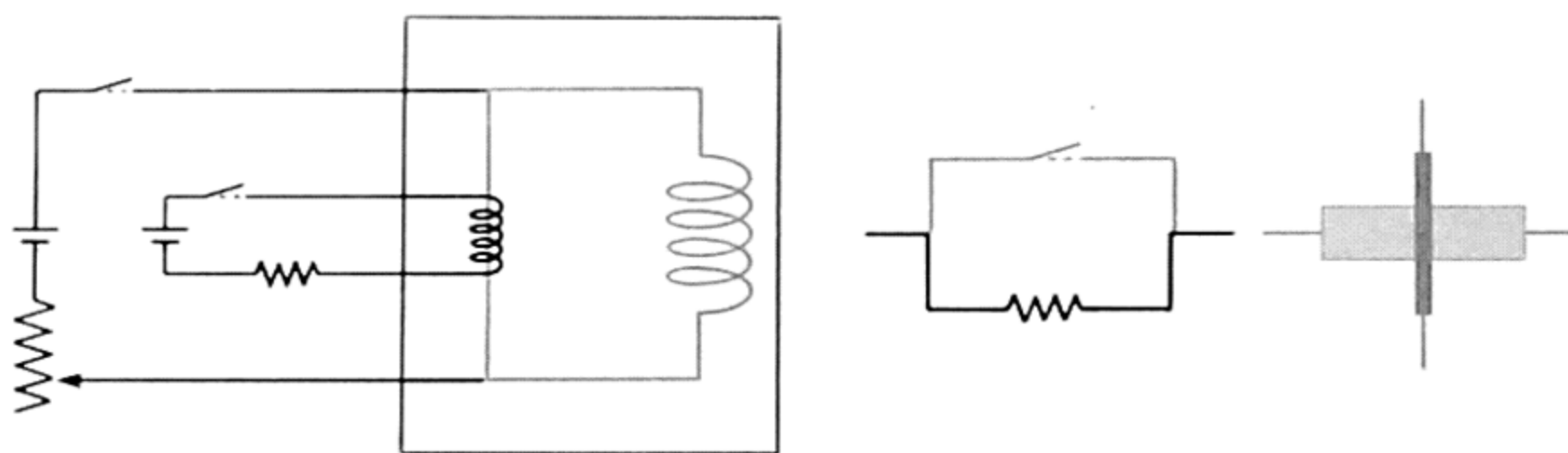
insulate against magnetic flux is applied in design of a bearing that has extremely low friction. As the superconductive disk descends under a load (*arrow*), it compresses the flux lines of the superconductive coil. Their increased density repels the disk more forcefully.



make atoms visible. Illustration at left shows unwanted stray lines of magnetic flux (*black lines*), which are among factors that limit ability of conventional lens to focus image-carrying electron



beam. Superconductive shields (*colored areas*) on the lens at right eliminate stray flux lines and shape the magnetic field, thereby improving its ability to focus the electrons and form a sharp image.



FLUX TRAP (left) is basis of several superconductive devices. Superconductive circuit (color) in low-temperature environment (gray) conducts current from source at left. Small heating coil keeps shunt (light color) resistive when switch is closed; when switch is open, coil cools and shunt becomes superconductive. Opening switch in nonsuperconducting part of superconductive circuit then cuts circuit from power source, trapping current and its flux indefinitely. Middle drawing shows analogy between resistive and superconductive states of shunt. When resistive, it is in effect an open switch; when superconductive, it is a closed one. Drawing at right shows a printed cryotron. When flux of vertical element exceeds the critical field strength of the horizontal element, the latter becomes resistive.

Superconductive amplifiers comprise another branch of the family of devices that derive from the basic magnetic bearing. In the bearing the vertical motion of the disk acts on the magnetic flux of the coil and converts its steady, or direct, current to a changing one. In the case of the amplifier the flux, instead

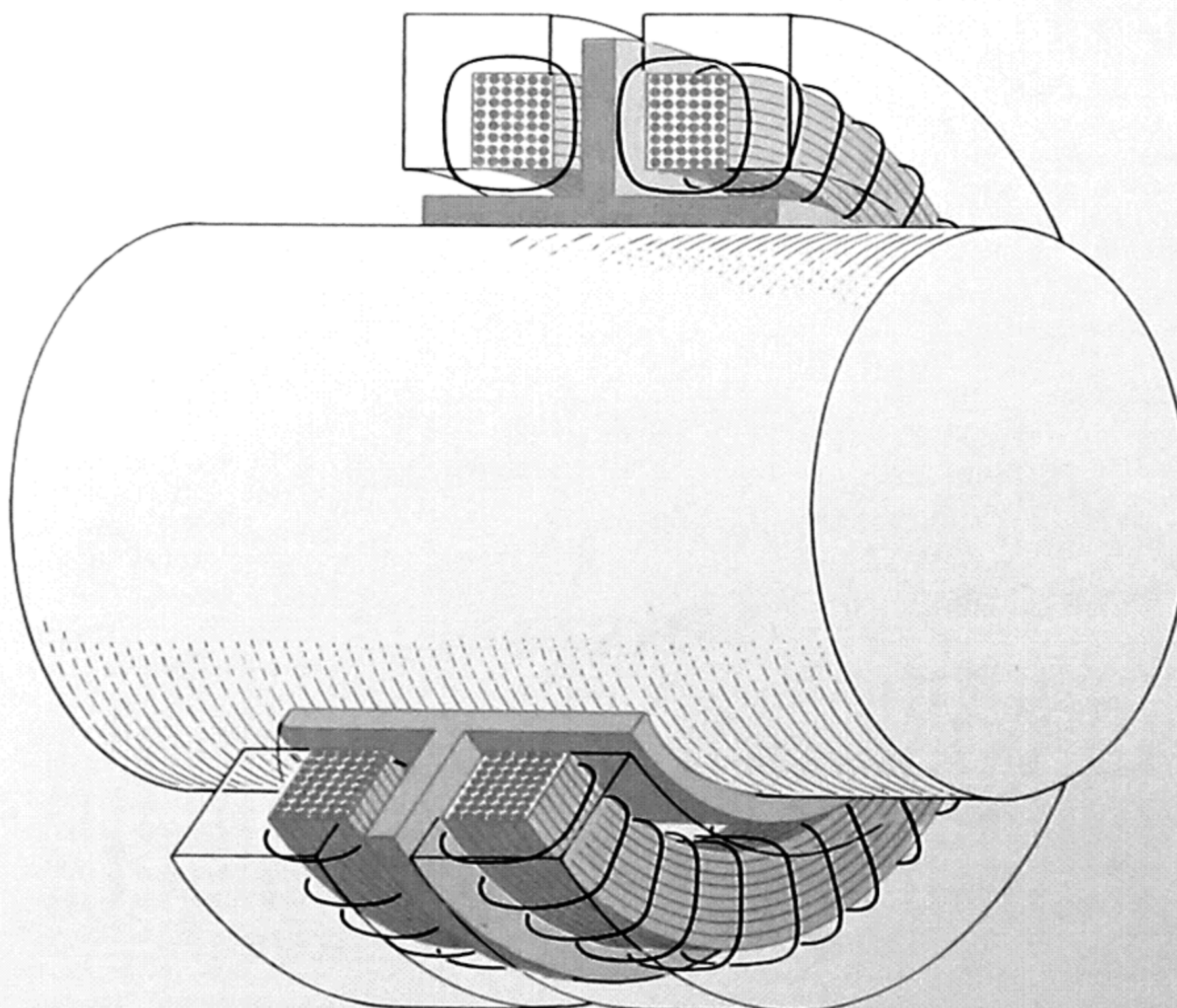
of being trapped, will be produced by a current fed in from a signal source. The up-and-down motion of the disk will now vary the density of the flux. With an insulated second coil wound around the first one the changing flux will induce an alternating voltage in the second coil. As the current in the first coil is in-

creased or decreased the output voltage of the second coil is increased or decreased proportionately. Here is the working principle of an amplifier that would convert direct current to alternating current. Since the superconductive coils offer no resistance to electrical current, regardless of the number of turns of wire in their construction, the amplifier would have infinitely high gain, or amplification.

To increase the operating frequency of such an amplifier, the disk can be segmented like a Maltese cross and, instead of moving vertically, can be made to rotate between four opposed pairs of fixed superconductive coils [see top illustration on page 590]. Flux of an encircling direct-current coil would be alternately repelled and passed by the blades and spaces of the rotating disk. The intermittent flux would induce an alternating current in the coils with a frequency dependent upon the speed at which the disk rotates. Such an amplifier would have zero drift, that is, perfect stability; and since it is always superconductive it would also be noiseless.

The responsiveness of superconductors to small changes in temperature and field strength qualifies them for service in high-precision switching and sensing devices. A superconductive heat-detector will operate in the narrow area within which resistance very rapidly declines to zero [broken line in top graph on page 585]. In this area very slight changes of temperature cause sharply defined changes in the resistance of the superconductor. The change in resistance thus provides an accurate measurement of change in temperature.

Superconductors are as sensitive to changes in magnetic-field strength as they are to changes of temperature. The cryotron, a relatively simple switching device, operates on the principle that a superconductor is rendered resistive by a magnetic field that exceeds its critical field strength. One element of a cryotron, a superconductive coil having a high critical field strength, surrounds the other element, a wire having a lower critical field strength. A slight increase in the coil current, and hence in its flux, renders the wire resistive; a decrease of coil current weakens its flux and allows the wire to become superconductive. The superconductive-resistive states of the cryotron correspond to the binary "on" and "off" states of ordinary computer elements. Since superconductivity is a surface phenomenon, very small cryotrons can be made with printed-circuit techniques. One element simply



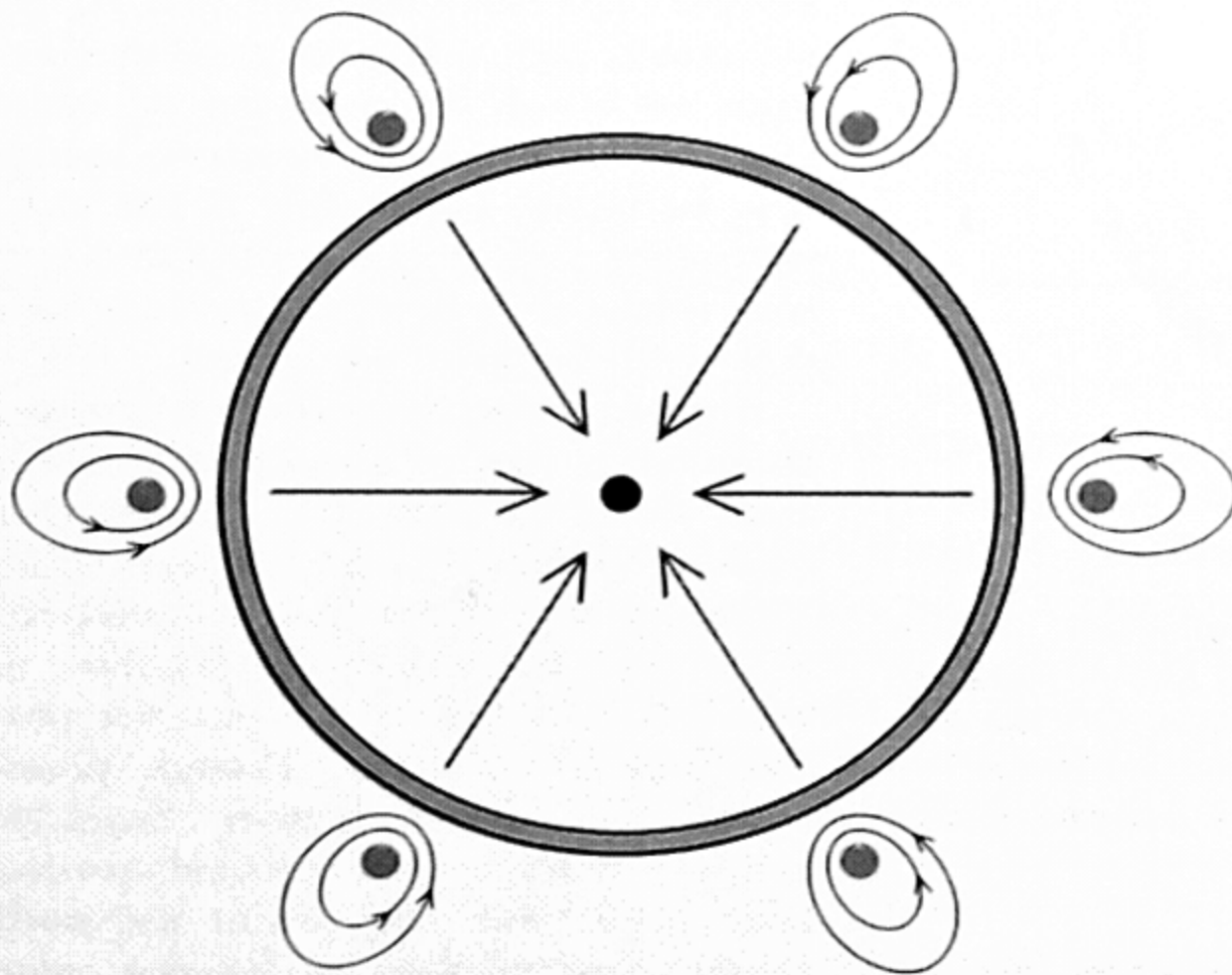
SUPERCONDUCTIVE BEARING supports a rotating shaft without friction. The T-shaped structure affixed to the shaft is superconductive. It rests on opposed magnetic fields (black loops), which are generated by concentric but noncontiguous coils, by repelling them.

intersects the other in the circuit [see top illustration on opposite page]. Printed cryotrons could reduce the bulk and the power requirements of present computers and enhance their reliability and versatility.

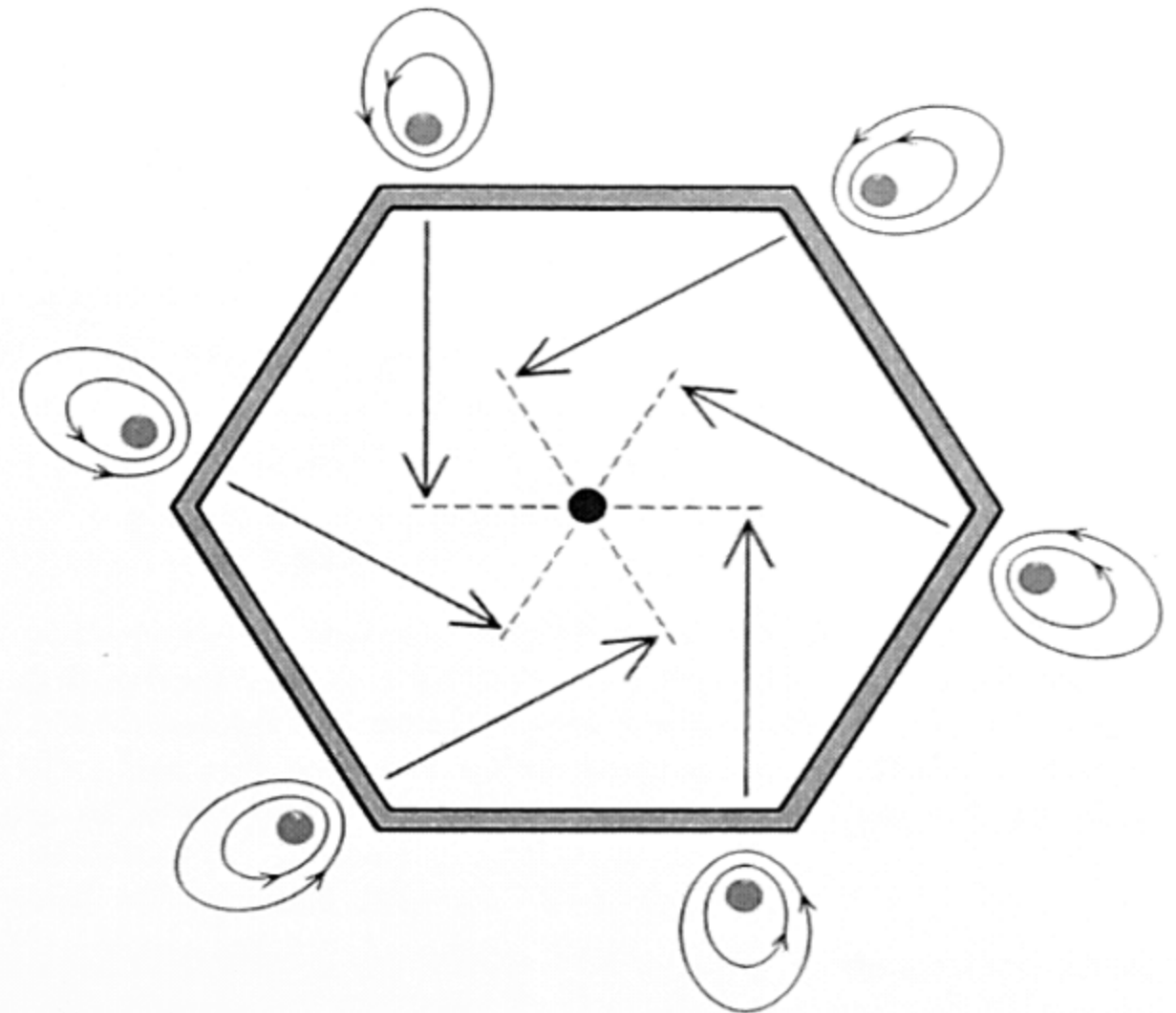
In the more distant future superconductivity promises improvements for special power-generating and transmis-

sion equipment. If the hollow conductors in the stators of a generator could be cooled sufficiently to make them superconductive, the generator could produce much more power. Liquid helium seems prohibitively expensive for this use. But materials might be found that become superconductive at the higher temperatures of liquid hydrogen (20.4

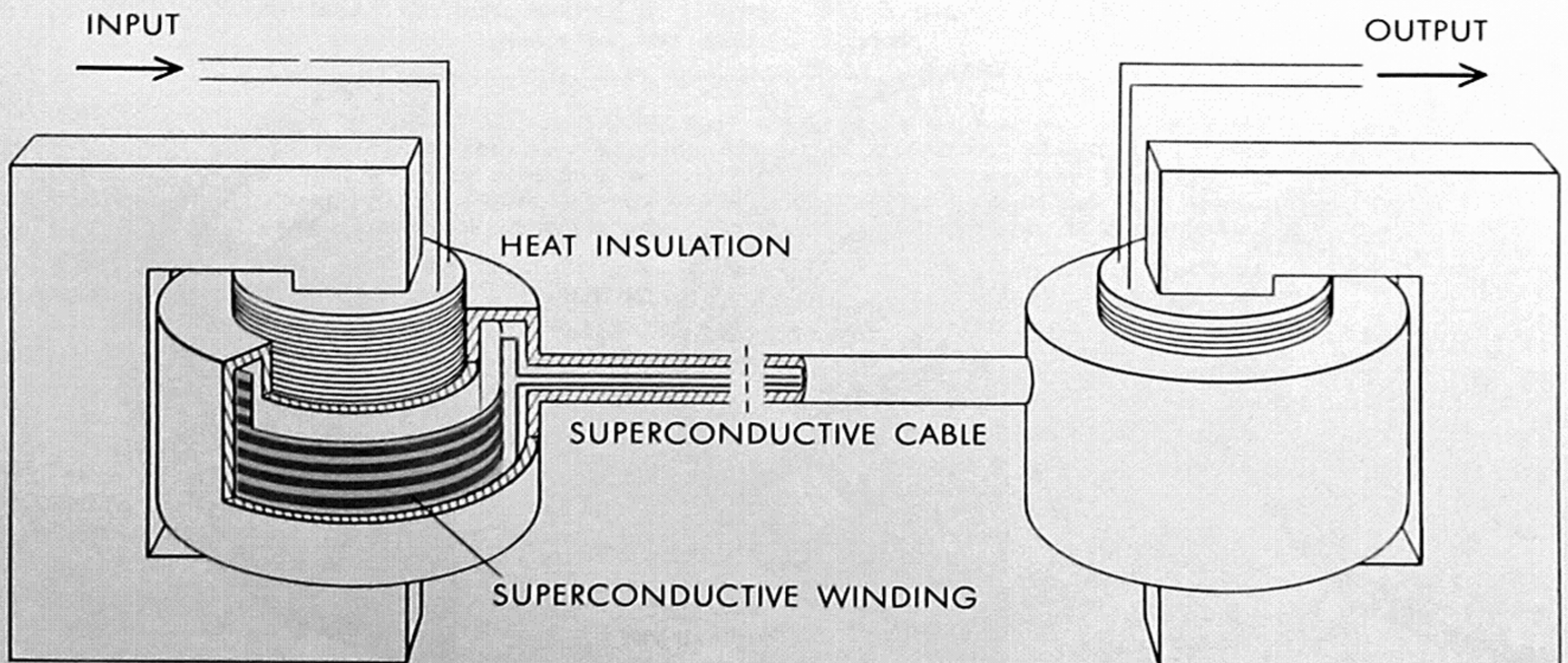
degrees) or, better still, of liquid nitrogen (77.3 degrees). Due to heat transfer through the driving shaft it does not seem possible to cool the rotor down to superconductive temperatures. The large currents produced by superconductive generators might be carried great distances with low losses to resistance by means of coaxial transmission



ROTOR OF SUPERCONDUCTIVE MOTOR would not rotate if it had conventional circular cross-section (*left*), because forces (*arrows*) exerted by magnetic flux (*ellipses*) of stator coil (*colored*

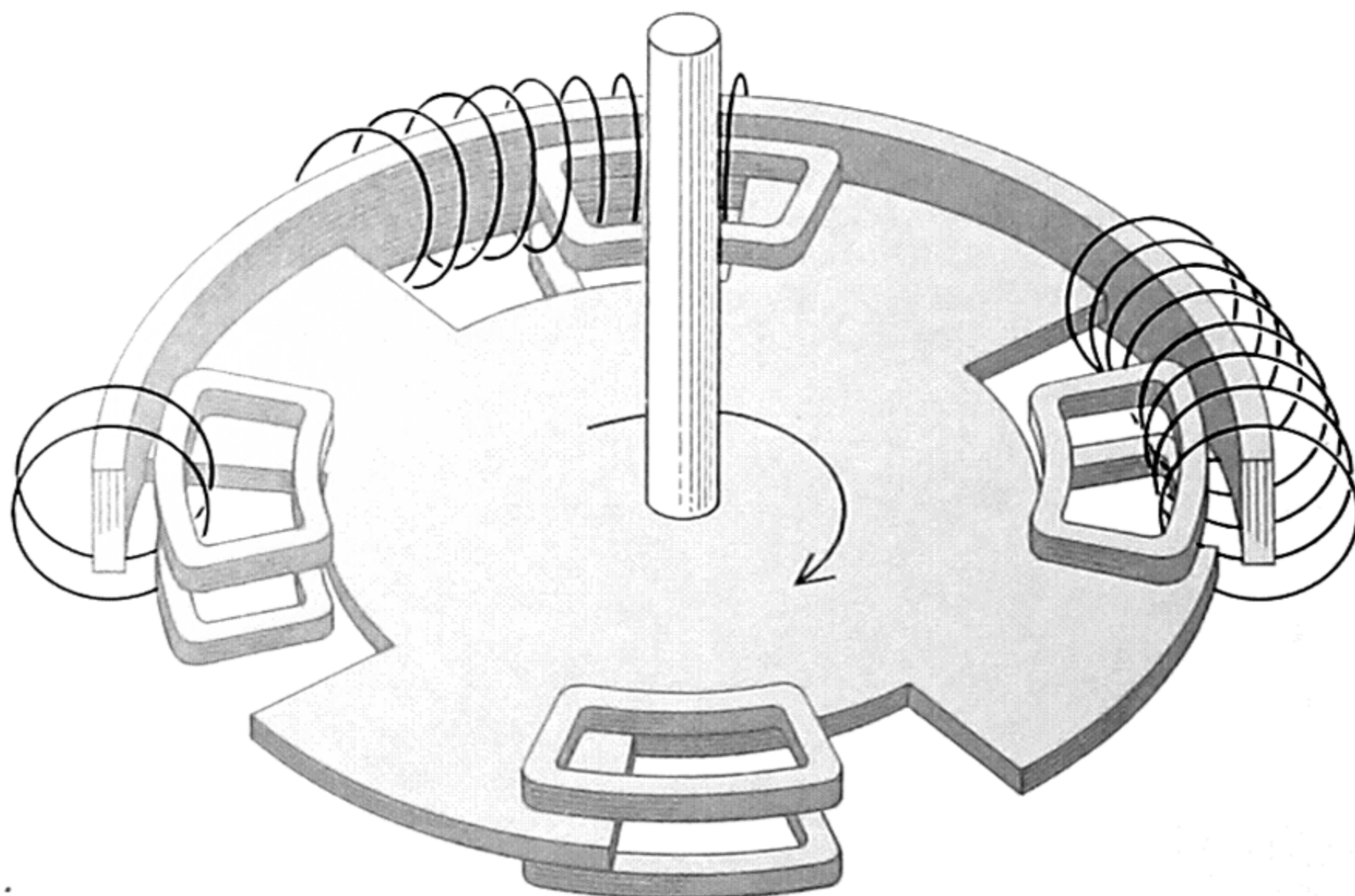


dots), being perpendicular to rotor surface, could not develop torque. Rotor with polygonal cross-section (*right*) would in effect have lever arms (*broken lines*) that would introduce torque.



PRINTED TRANSFORMER COILS, made feasible by superconductive metals, could replace the massive coils of present transformers. Printed on an insulating base, the superconductive coils

(*shown in color*) would be immersed in a low-temperature bath of liquid hydrogen or nitrogen and insulated against ambient heat. Superconductive cables would be used to complete power circuits.



SUPERCONDUCTIVE AMPLIFIER converts direct current, in fixed outer coil, to alternating current in fixed coil pairs. As the superconductive disk rotates, its blades block the lines of magnetic flux (*black loops*) generated by outer coil, and its spaces pass the flux. Alternate blocking and passing of the constant flux make the current induced in the coil pairs an alternating one. The shaft, which rotates the disk, is powered by external source.

lines whose center conductors would be superconductive.

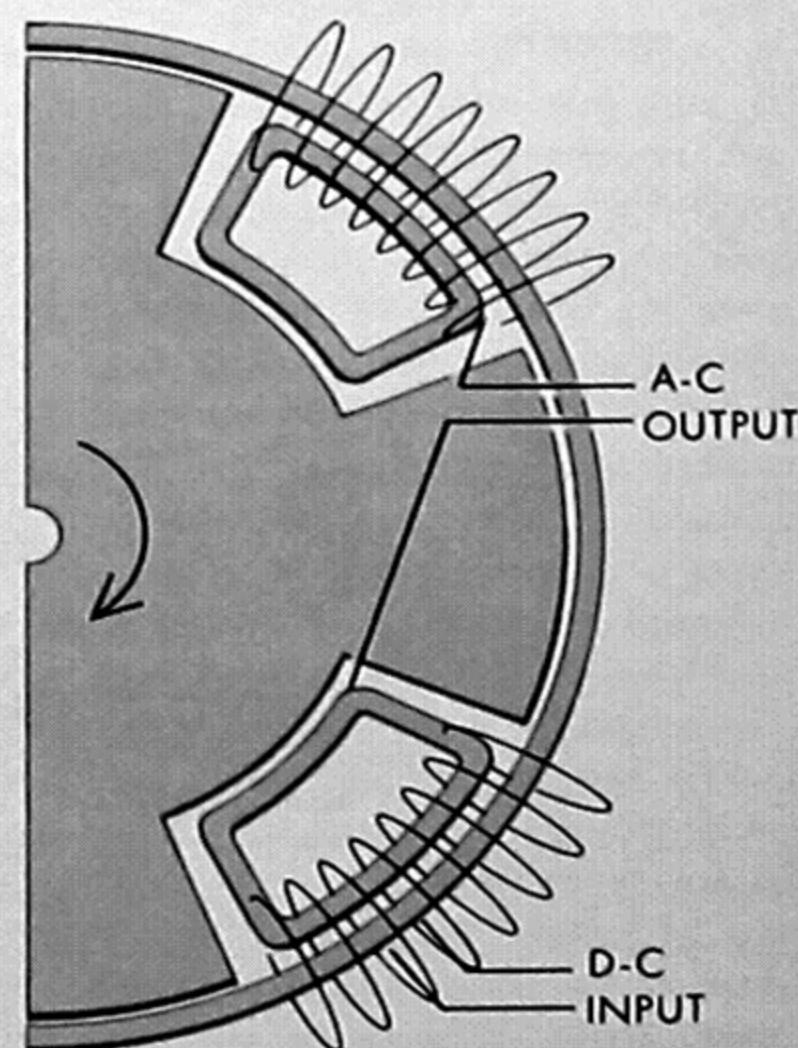
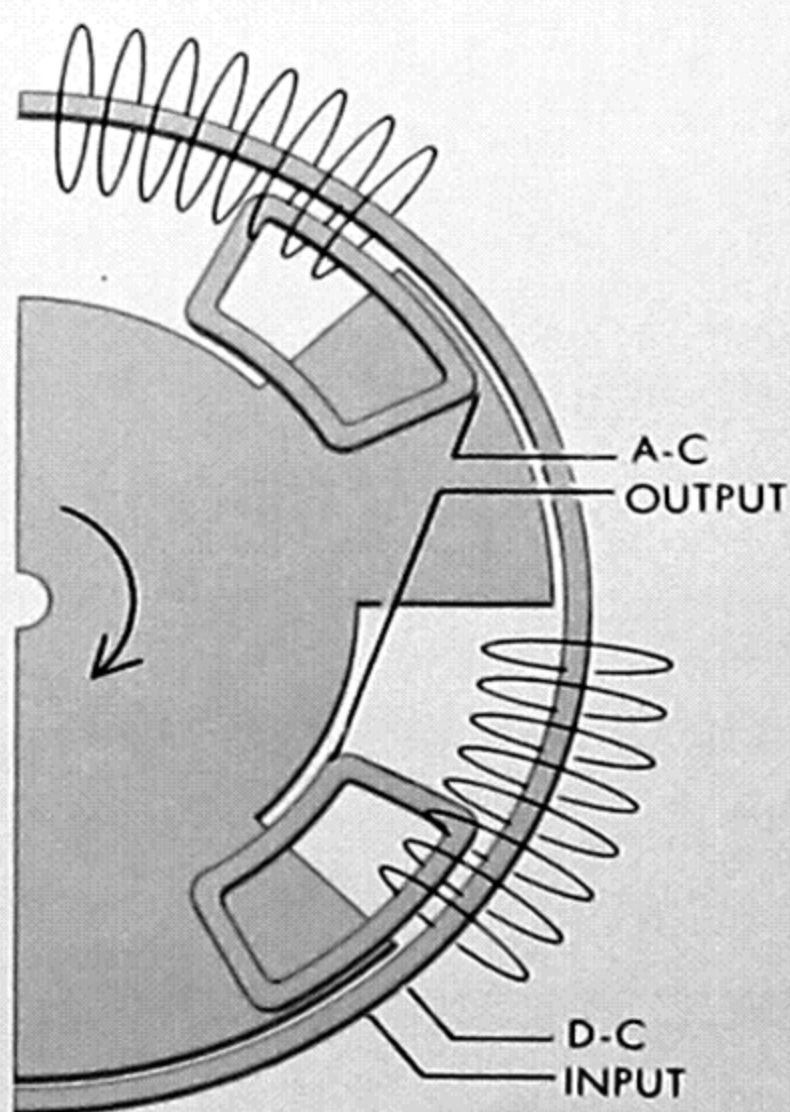
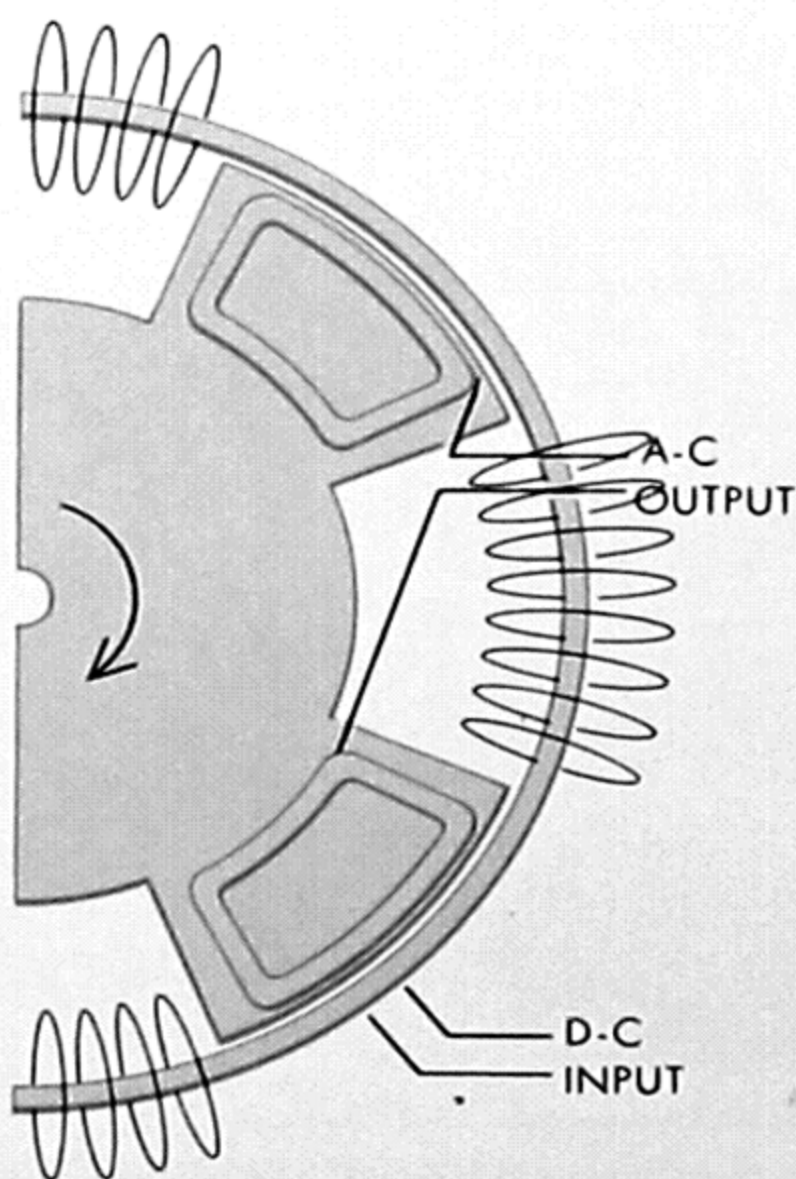
Such superconductive cables could feed power to superconductive transformers. But the usual power-loss in a transformer can be reduced by making only one of its windings superconductive. This would provide a further sig-

nificant advantage. Instead of massive quantities of metal in huge coils, the superconductive winding would consist of a cylinder bearing a printed coil [see *bottom illustration on preceding page*]. In installations requiring two transformers, one coil of each transformer might be installed in a low-temperature envi-

ronment; the other coils would be outside this environment.

The particle accelerators employed in nuclear physics might gain immediate advantage from progress in the technology of superconductivity. Their magnetic coils consume great quantities of power, much of which is lost to resistance. As superconductive alloys with higher critical field strengths are developed and fashioned into coils for accelerators, they will quite possibly eliminate all losses. By applying the principle of the flux trap, magnetic energy may be stored indefinitely in the magnetic coils and, under appropriate control, released in short pulses as is now done with banks of capacitors.

Applied superconductivity is a young technology, now at a stage of development similar to that of electricity after Michael Faraday made his discoveries a century ago. Much that is envisioned is not yet technologically feasible, or where it is feasible, it is not yet economically competitive. Though prototypes of superconductive bearings, motors, switches and resonant cavities have been built, they await improved refrigerating systems to become more attractive. Meanwhile more research should be devoted to attaining higher critical field strengths and higher transition temperatures. With adequate enterprise along both lines, the meeting of the two technologies may occur sooner than we can now foresee.



OPERATION OF SUPERCONDUCTIVE AMPLIFIER is based upon rotation of superconductive disk. At left spaces of disk pass constant flux (*black loops*) set up by peripheral coil. As disk ro-

tates, flux reaches small coils (*drawings in middle and at right*), where it induces a current. Induced current alternates between positive and negative values as the flux is blocked and passed.

The Author

THEODORE A. BUCHHOLD is a consulting electrical engineer in the General Engineering Laboratory of the General Electric Company in Schenectady, N.Y. A native of Germany, he acquired his master's and doctor's degrees in electrical engineering at the Darmstadt Technical University. After working in private industry he became a professor at Darmstadt in 1934. During World War II Buchhold developed a very accurate accelerometer for V-2 rocket guidance and pioneered in work on magnetic amplifiers which were used in the V-2. In 1946 he joined the Guided Missile Development Division at Fort Bliss, Tex., as chief of the control and guidance branch, which was later moved to the Redstone Arsenal in Huntsville, Ala. He developed the basic concept of a new inertial-guidance system for the Red-

stone missile. Buchhold became consultant to the vice president of engineering of the Ford Instrument Company in New York City in 1954, and a year later joined the G.E. laboratory.

Bibliography

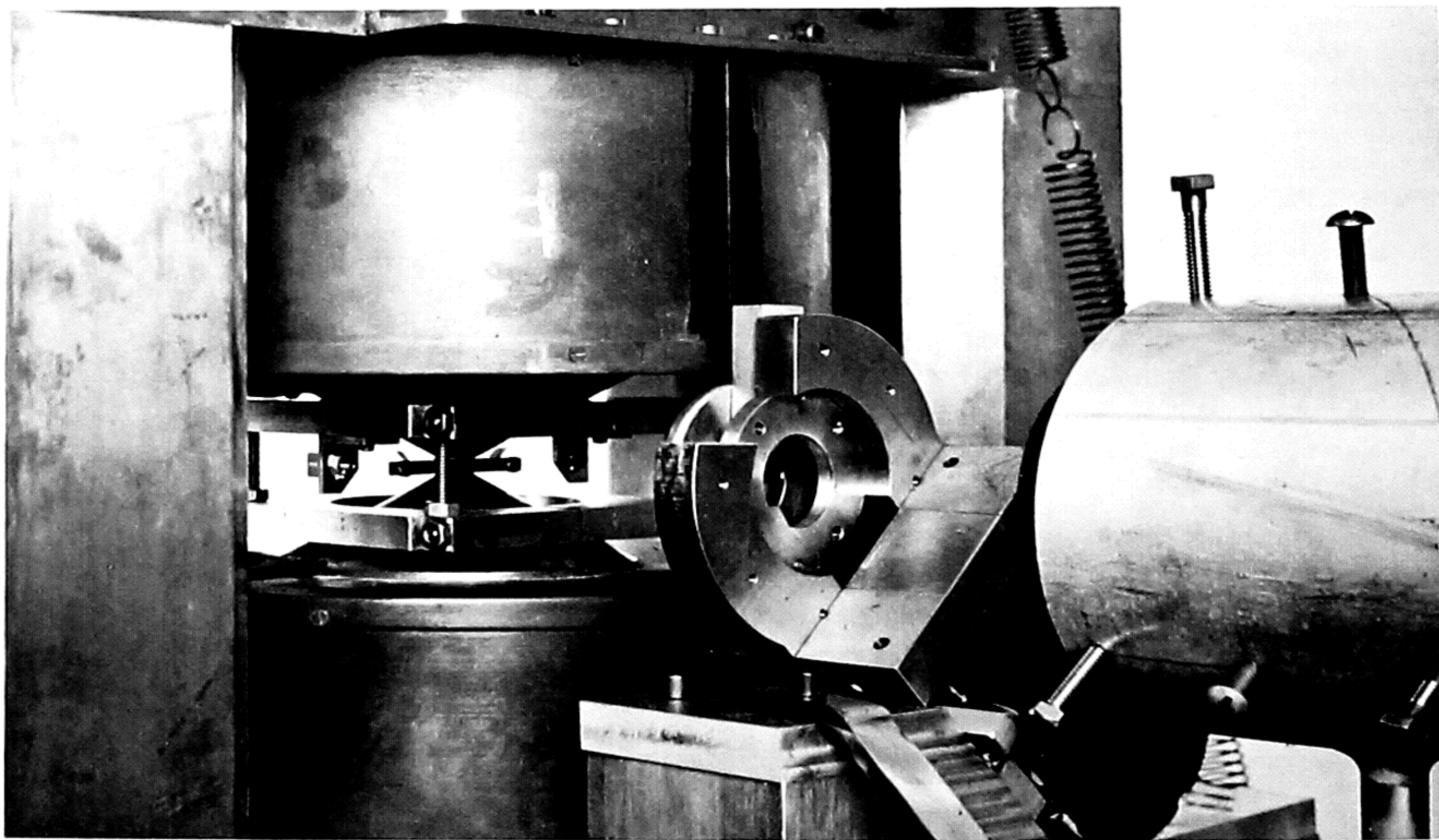
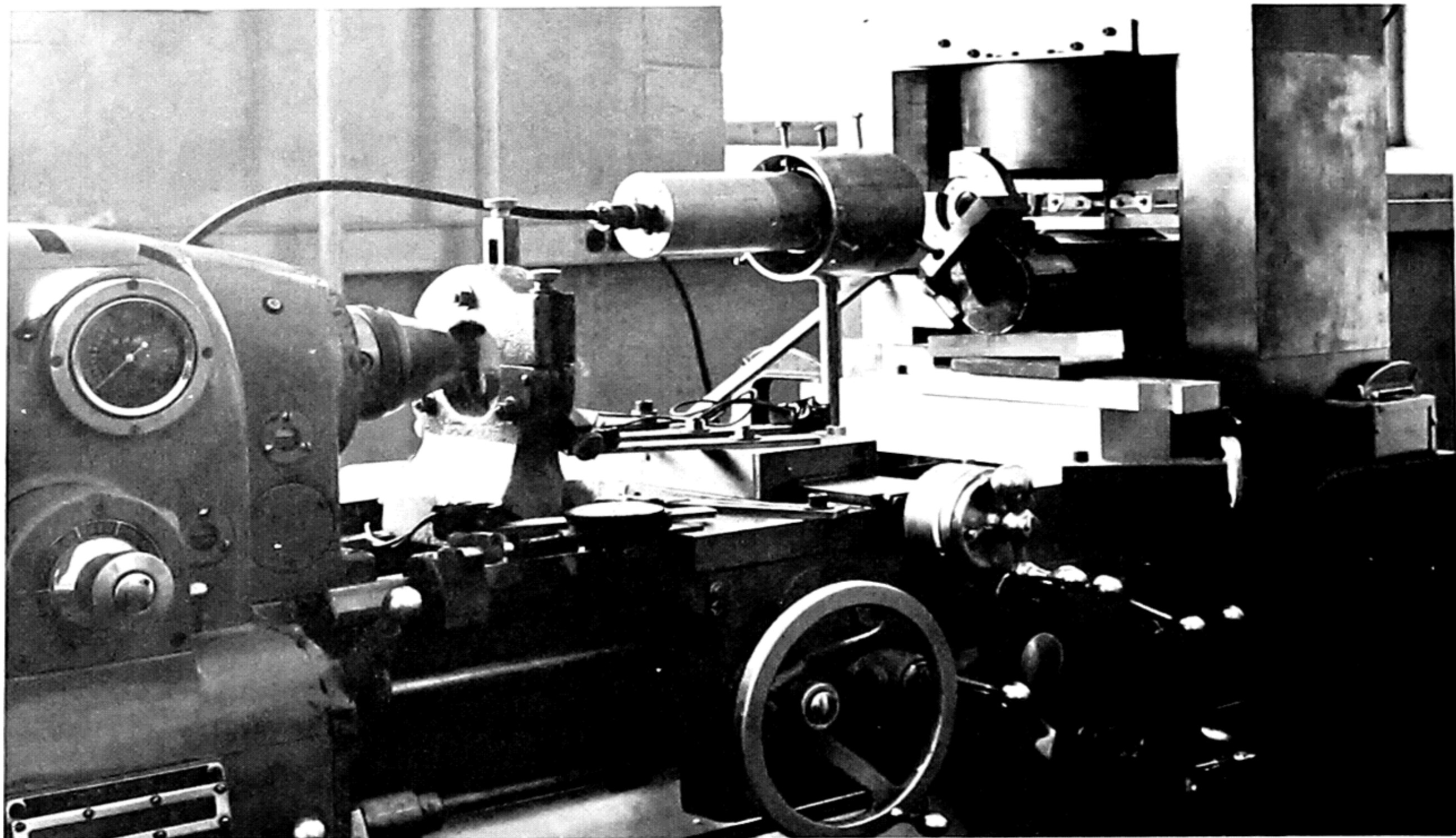
FORCES ACTING ON SUPERCONDUCTORS IN MAGNETIC FIELDS. I. Simon in *Journal of Applied Physics*, Vol. 24, No. 1, pages 19-24; January, 1953.

LOW TEMPERATURE PHYSICS I AND II. *Encyclopedia of Physics*, Vol. 14 and 15; 1956.

SUPERCONDUCTIVITY. David Shoenberg. Cambridge University Press, 1952.

THEORY OF SUPERCONDUCTIVITY. J. Bardeen, L. N. Cooper and J. R. Schrieffer in *The Physical Review*, Vol. 108, No. 5, pages 1,175-1,206; December 1, 1957.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.



RESONANCE-ABSORPTION EXPERIMENT was photographed at the Argonne National Laboratory. Assembly of iron-57 source of gamma radiation together with absorber and detector can be seen at upper right in top picture, mounted on a lathe that moves the

absorber. In the close-up of the assembly (*bottom*) the source is the flat plate between tapering magnetic poles (*left*). Absorber is mounted in circular disk (*center*); detector is in cylinder (*right*). Magnet is used to measure "hyperfine" splitting of absorption line.

THE MÖSSBAUER EFFECT

by Sergio De Benedetti

A German physicist's discovery is the basis of a "nuclear clock" of unprecedented accuracy. Such a device should make possible the first conclusive test of the general theory of relativity.

According to a very old and very wise myth, Zeus, the chief deity of Olympus, was the son of Chronos, the god of time. The legend demonstrates the Greeks' deep intuitive understanding of the natural world; they realized that time is the first of all mysteries, the most fundamental of all concepts.

Just recently there has fallen into man's hands an almost unimaginably sensitive technique for measuring time and penetrating some of its long-kept secrets. Thus Chronos is suffering the fate of his Olympian descendants, whose places were usurped by laws of nature as Greek intuition gave way to the quantitative reasoning and experimental methods of our scientific age.

We begin the account of this development in the Attic spirit, with an embellishment of the ancient legend. Imagine that at the first Olympic games, a few hundred years before the time of Christ, Chronos himself awarded a prize to the winner of the marathon. Of course the prize was a watch, but a watch of perfect accuracy such as only the god of time could give. We may well suppose that this matchless timepiece was reverently handed down from father to son. Designed like a modern watch, it ticked away through the centuries exactly five times per second, measuring the absolute and immutable flow of time.

Then came Albert Einstein to proclaim the relativity of time. His special theory of relativity showed that the rate at which clocks run depends on their relative motion. If a fast runner carried Chronos's gift, it would lose time at the rate of about a hundredth of a tick in a million million. Not a very serious loss, but in principle an affront to the dignity of the god. Furthermore, according to the general theory of relativity, the

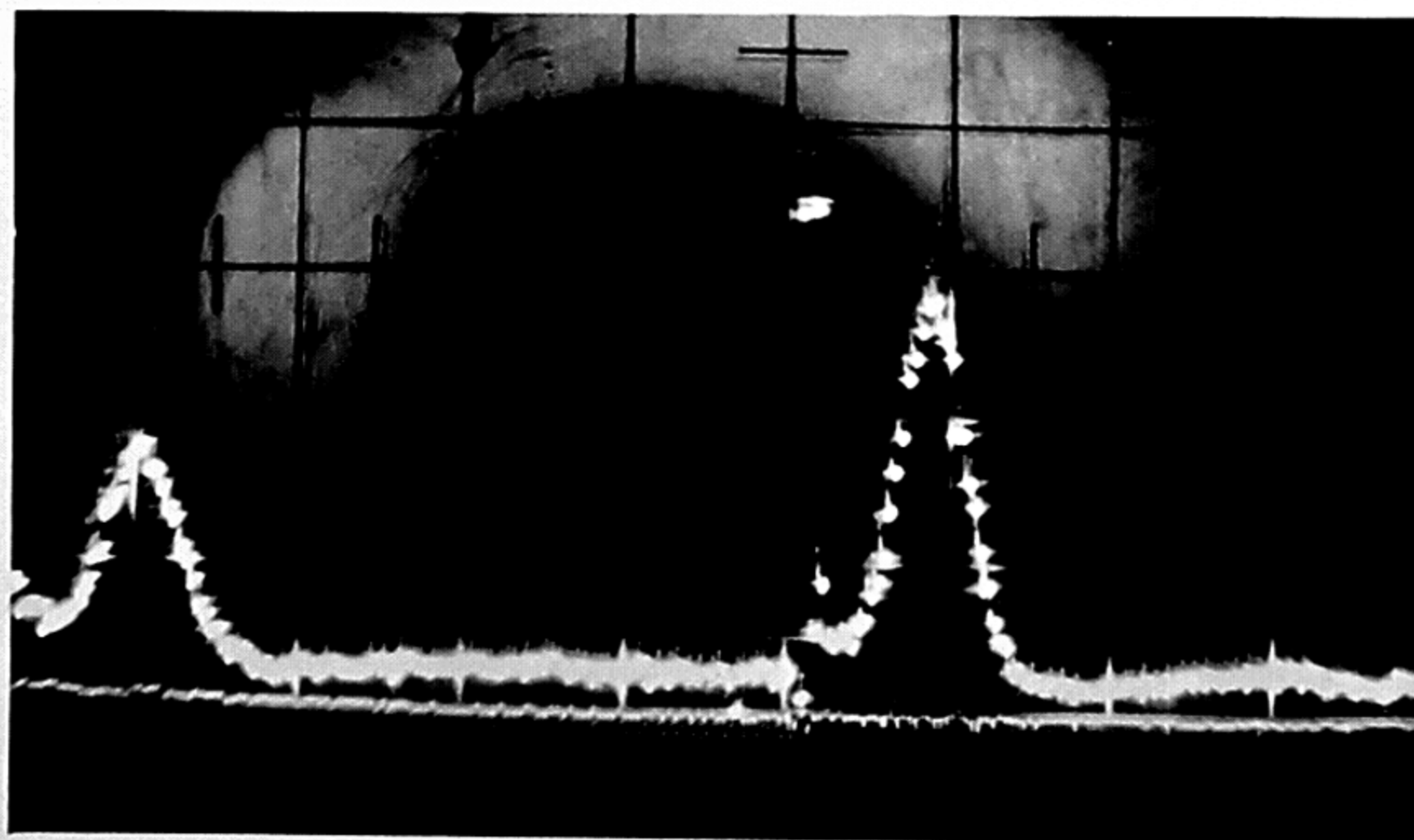
watch would speed up by about the same amount if it were kept on the top floor of a building a few stories high rather than on the ground floor, because of the difference in the gravitational potential energy.

Still, as an anthropomorphic god, Chronos could afford to shrug off relativity. Perhaps subatomic particles, moving at nearly the speed of light, would demonstrate the contraction of time with velocity; perhaps the gravitational effect could be perceived in the red-shift of light from the dense companion of Sirius. But on earth, for distances and speeds on the human scale, the effects would be undetectable. And so they were until a few months ago, when the German physicist R. L. Mössbauer published a paper that opened the way to a meas-

urement of the loss of a hundredth of a tick in a million million!

To understand the significance of Mössbauer's contribution, consider the requirements of an instrument to measure time. First of all, there must be a periodic device—something that undergoes a repetitious, or cyclical, motion always in the same length of time—like a pendulum or a balance wheel. But for extremely accurate studies mere regularity is not enough. There must be a substantial number of repetitions in a reasonable time. The faster the ticking and the greater the number of uninterrupted ticks, the better.

We can see why this is so if we suppose that we really have Chronos's watch, and that its perfect balance-wheel



TRANSMISSION OF GAMMA RAYS in absorption experiment is recorded as vertical deflection on oscilloscope. Peak at left shows transmission when source and absorber are at rest and therefore in resonance. Peak at right shows increased transmission when source and absorber are "detuned" by moving absorber at the rate of 1.17 millimeters per second.

is the fastest pendulum device available to us. All slower processes could be measured in terms of its one-fifth-of-a-second ticks. But what if we made a second watch as closely like it as possible and set out to compare the two? Assume that this second device actually runs slow at the rate of one tick in 10^{12} (a million million). Assume also that we can detect that the two watches are not synchronized when they differ by as little as, say, a 10th of a tick. If they start out exactly together, they must tick 10^{11} times before we can tell that one is running slower. At the rate of five ticks per second this would take about a millennium.

In the past few years physicists have found much more rapid and accurate pendulums. Using the vibrations of certain molecules or atoms, they have built atomic clocks that tick several thousand million times a second [see "Atomic Clocks," by Harold Lyons; SCIENTIFIC AMERICAN Offprint 225]. But nature has provided a still better timekeeper in the nucleus of the atom.

In classical terms (by which is meant not the ancient Greeks and their gods, but simply a description that does not involve quantum ideas) we may think of a nucleus as a spherical body with a uniformly distributed positive electric charge. The sphere is not rigid, but behaves like a liquid drop. If the nucleus is excited, the sphere vibrates like a drop of water or mercury. When

it vibrates, the nucleus radiates electromagnetic waves, which we call gamma rays. The emission of the waves requires energy; the amplitude of their vibration gradually decreases in the same way that the vibrations of a violin string die out as their energy is carried away by sound.

In a typical case the frequency of the gamma-ray vibration may be 10^{18} (a million million million) oscillations per second. Since certain nuclei emit energy at a very slow rate, they keep on vibrating for a relatively long time, a time which is measured in seconds, days or even months. These long-lasting excited states of nuclei are called isomeric states, and their duration is usually expressed in terms of a half-life—the time required for the intensity of the emitted radiation to decrease by half.

An excited nucleus with a long half-life is a virtually perfect pendulum. All nuclei of the same kind are exactly alike; this is a great advantage over man-made pendulums, no two of which are ever really identical. Moreover, being protected from external influences by surrounding atomic electrons, a nucleus vibrates at a rate that is not affected by external influences such as temperature or chemical change. Finally, a nuclear pendulum is not damped by unpredictable frictional forces, and thus "ticks" with absolute regularity an enormous number of times without requiring to be pushed again. Even if the half-life

is only a millionth of a second, there will be 10^{12} vibrations before the oscillations are considerably damped.

A good pendulum is the heart of a good clock, but it is not the clock itself. There must also be a device, such as a dial with hands, that counts the swings of the pendulum and thus allows us to read the time. We cannot directly count the oscillations of a vibrating nucleus. What we can do, however, is to compare with great accuracy the frequencies of two nuclear pendulums.

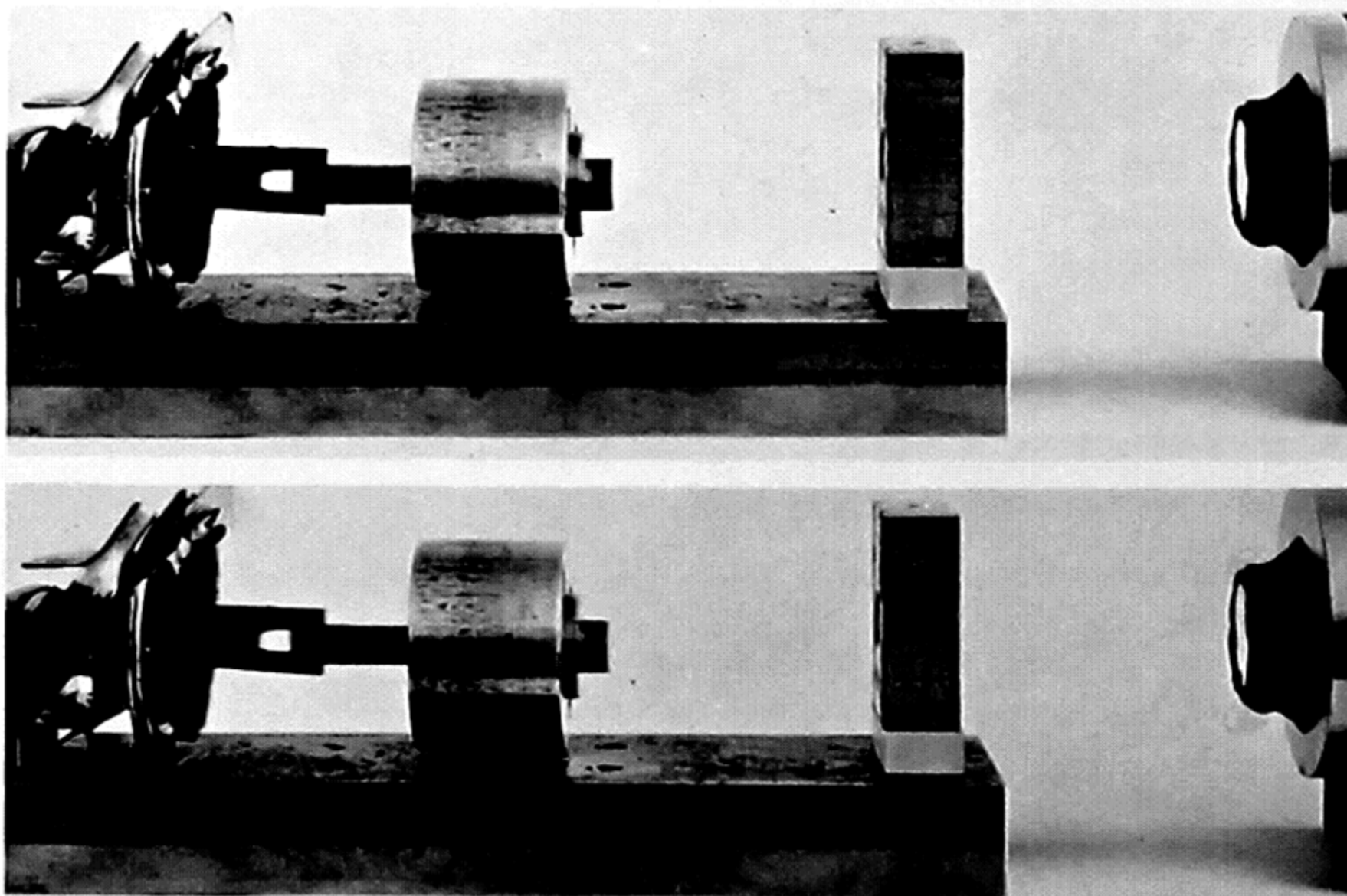
The method is based on the familiar observation that sound waves from a vibrating piano-string induce resonant vibrations in another string, initially at rest but tuned to the same note. In the same manner gamma rays emitted by an oscillating nucleus are absorbed by another nucleus of the same kind, and set it into vibration. When we observe this, we know that the two nuclei must have the same vibration rate, within very narrow limits of error.

What sets the limits? As our earlier example indicated, the accuracy that can be achieved in measuring a frequency increases with the number of oscillations involved. This means that a nucleus with a long half-life has an intrinsically better-defined frequency than a nucleus of short half-life. But resonance excitation requires that the frequencies of emitter and absorber be the same within their intrinsic accuracy. It follows that the "sharpness" of the resonance increases with the half-life of the excitation, which we have already deemed to be related to the quality of the clock.

In our typical case of 10^{12} spontaneous nuclear oscillations the resonance effect disappears if the frequency of the emitting nucleus differs from that of the absorber by as little as one part in 10^{12} . In the language of modern physics we say that the width of the resonance, or the line width, is one part in 10^{12} .

At present the most convenient material to use in resonance experiments is the iron isotope of mass 57. One starts with a radioactive source of the isotope cobalt 57, which is commercially available and has a convenient half-life of 280 days. As the cobalt-57 nuclei decay at this leisurely rate, they change into excited nuclei of iron 57. The iron nuclei vibrate at a frequency of 3×10^{18} oscillations per second, with a half-life of 10^{-7} second (a 10th of a millionth of a second). Thus an iron-57 nucleus emits roughly 10^{12} waves.

In the resonance measurement a narrow beam of the waves is aimed at an-



RELATIVE MOTION of source and absorber can also be produced by linking the source (*thin cylinder*) to a loudspeaker (*left*), which vibrates in response to an audio-frequency signal. Gamma rays pass through absorber (*center*) and into detector (*right*). Source appears sharp when speaker is still (*top*); slightly blurred when it vibrates (*bottom*).

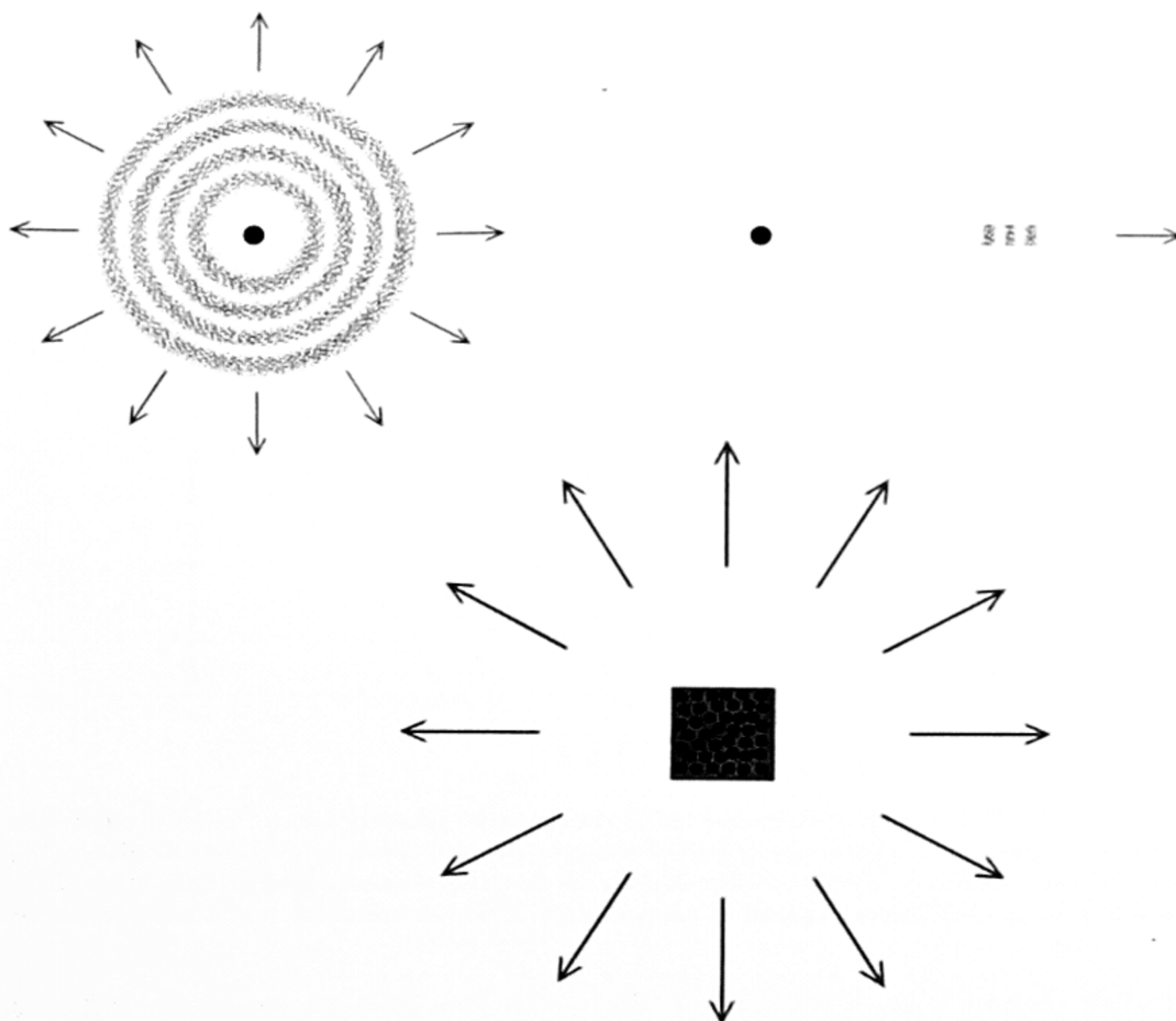
other piece of iron containing stable nuclei of iron 57 [see bottom illustration on this page]. At resonance these nuclei strongly absorb energy from the beam and reradiate it in all directions. A scintillation counter set up beyond the absorber in the line of the original beam records a sharp drop when the absorber nuclei are in resonance with the beam. At the same time a counter located near the absorber at right angles to the beam direction shows an increase in energy due to the scattered radiation emitted by the absorber nuclei. The experiment is not at all difficult to perform. If one uses pure iron 57 for the absorber, the difference in the transmitted energy recorded by the first counter at resonance and out of resonance is as large as 50 per cent.

The theory of nuclear resonance goes back several years, and, as we have remarked, the crucial experiment is a comparatively easy one. Why was it not done earlier? The reason is that certain secondary effects made it quite difficult to observe nuclear resonance before the work of Mössbauer.

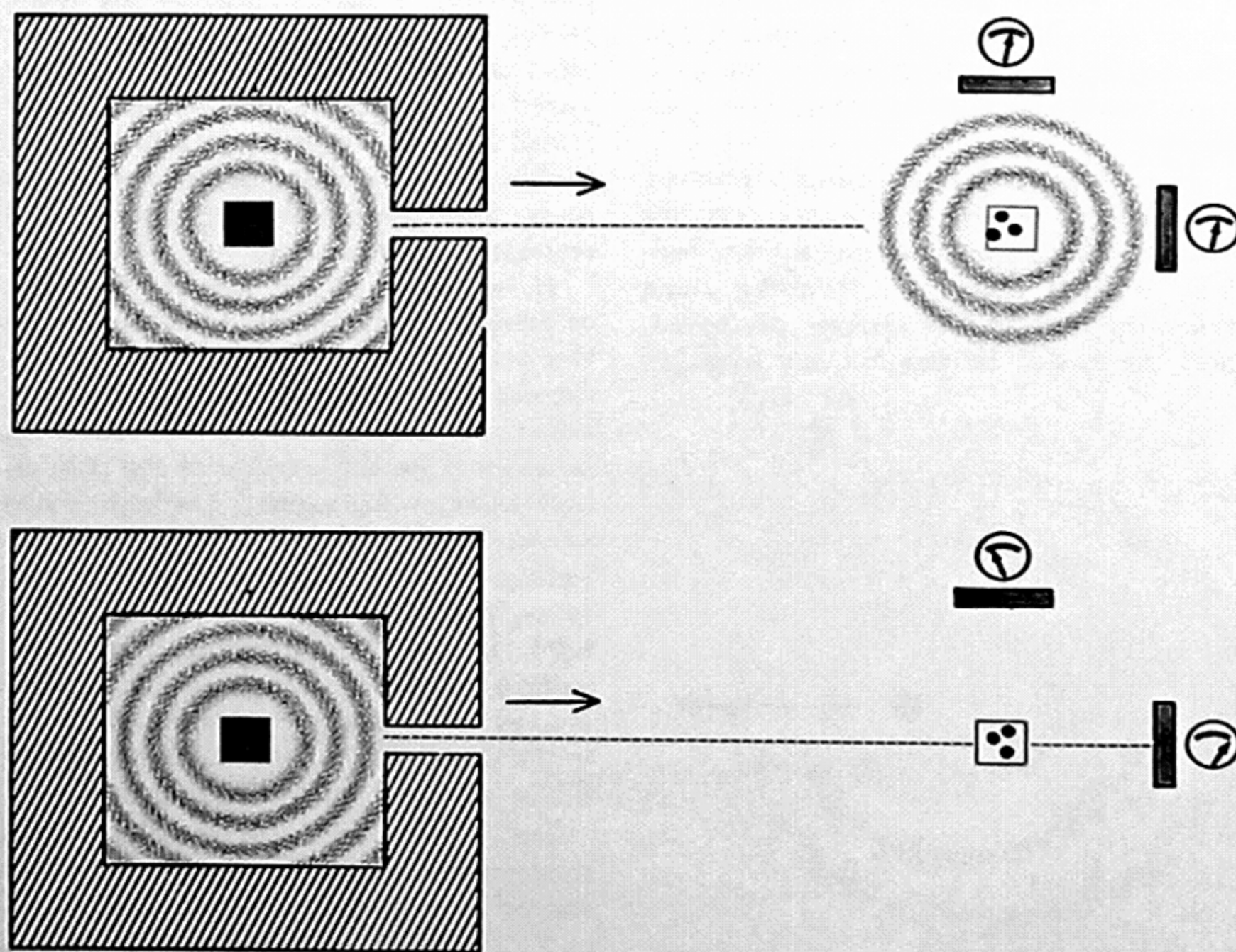
To explain these effects we shall have to abandon the oversimplified, classical picture of nuclear radiation and consider some of its quantum aspects. On this view an excited nucleus does not lose its energy gradually by radiating a continuous wave in all directions. Instead, at an unpredictable moment, it drops abruptly from its excited state to its stable, or "ground," state, emitting a "packet" of radiation in a single, unpredictable direction. This packet, or photon, carries an amount of energy equal to the difference in energy between the excited and ground states of the emitting nucleus. The absorption process is just the reverse. A nucleus in the ground state absorbs a photon and is raised to the excited state. Some time later it reverts to the ground state and reradiates the energy as a new photon.

As Einstein showed many years ago, the energy carried by a photon is proportional to its frequency. Thus in a certain sense the words "energy" and "frequency" can be used interchangeably in quantum-mechanical language. For example, the poor definition in frequency due to a short half-life now becomes a poor definition in energy, and the physicist speaks of the width of an energy line.

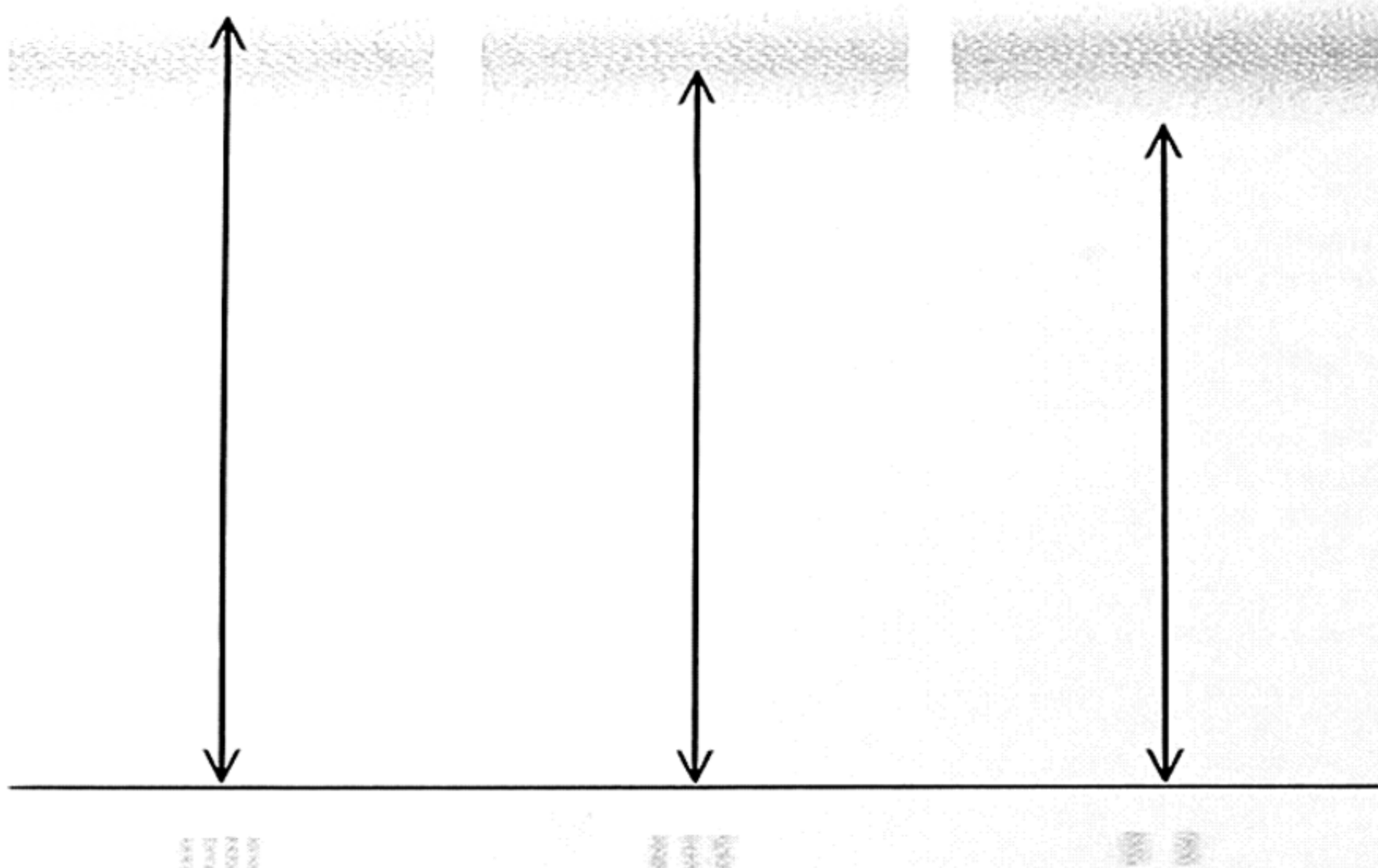
In quantum-mechanical parlance resonance occurs only if the energy lines of the source and absorber overlap within their widths. In the case of iron 57 the photon has an energy of about



NUCLEAR RADIATION is illustrated schematically in "classical" terms (*top left*) and quantum terms (*top right*). Dot represents an isolated nucleus; circles represent spherical waves in cross section. Rectangular wave-packet at right represents a photon. On either picture an assembly of many nuclei would emit radiation equally in all directions (*bottom*).



RESONANCE is demonstrated by directing a beam of gamma rays from excited nuclei (*dot at left*) at target nuclei (*dot at right*). Detectors (*rectangles*) in line with beam measure transmission; those at right angles to beam, scattering. Meter dials represent schematically the values of transmission and scattering at resonance (*top*) and out of resonance (*bottom*).



QUANTUM TRANSITIONS between high energy-level (*shaded bands*) of an excited state and low energy-level (*horizontal line*) of the ground state result in emission or absorption of photons (*bottom*). Photons vary slightly in frequency (and spectral lines have finite widths) because of energy spread of excited state. The average value is the most probable.

14,000 electron volts. If we think of this energy as represented by a line in a spectrum of energies, then the width of the line is determined by the half-life of the excited state of iron 57. As we have mentioned, the half-life is about 10^{-7} second, which gives a line width of about 10^{-8} electron volt. This means that the uncertainty or fuzziness in the photons' energy is approximately one part in 10^{12} .

Now when a nucleus emits a photon, the nucleus recoils in the opposite direction, like a gun that has fired a bullet. In both cases the recoiling mass takes up part of the energy produced, and the bullet or gamma ray emerges

with less than its maximum available energy. On the other hand, a nucleus that absorbs a gamma ray recoils too, and thus requires a photon with somewhat greater energy. As a result both emission and absorption lines are displaced. In the case of iron 57 the recoil energy is about 10^{-3} (one thousandth) electron volt. This is small indeed compared to the energy of a photon, but it is still 100,000 times larger than the line width. Thus the emission and absorption lines are completely separated, and resonance cannot occur.

If we think of the process in terms of frequency rather than energy, then the destruction of resonance can be understood as resulting from the Doppler effect. A moving source of radiation, whether it be the whistle of the proverbial train or an excited nucleus, emits waves that are packed together more tightly in the forward direction and more loosely behind [see illustration at left]. Hence the radiation from a recoiling nucleus, which is of course always emitted in the direction opposite to the recoil, has a lower frequency than if the nucleus were standing still.

But if the motion of recoil destroys resonance, an opposite, offsetting motion should re-establish it. And so it does, as P. B. Moon and A. Storruste of the University of Birmingham demonstrated in 1953. They mounted a source of excited mercury-nuclei on the arm of a centrifuge. When the arm was rotated

at the right speed, they observed resonance in a stationary mercury absorber. The motion required to offset the nuclear recoil in mercury was of the order of the speed of sound.

Moon showed how to compensate for the recoil. Mössbauer showed how to eliminate it. If we want to keep a gun from recoiling, so that the whole energy of the explosive charge will be imparted to the bullet, we must fasten the gun to the ground or to some heavy object. This is exactly what Mössbauer did to the nuclei; he fastened them to a piece of solid matter.

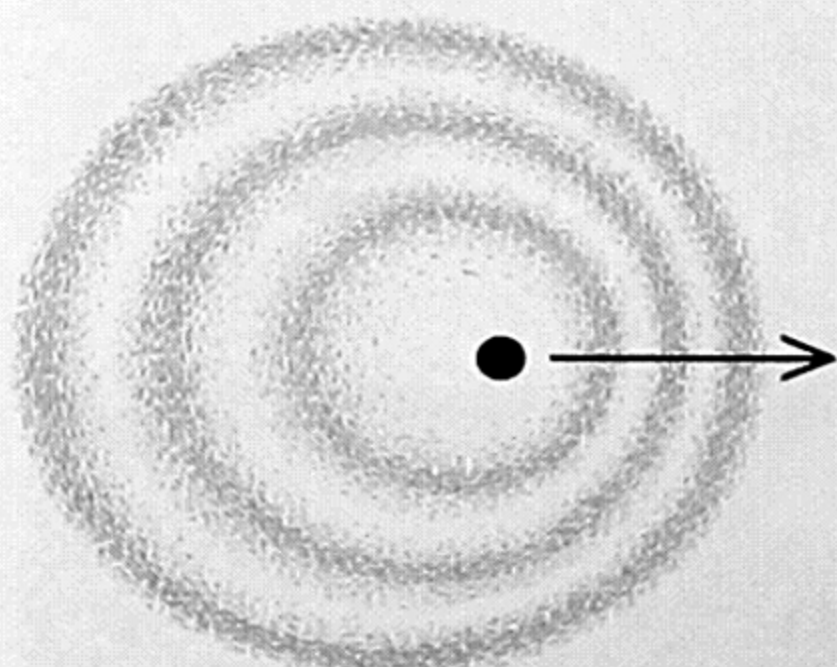
He realized that, under proper conditions, the forces that hold atoms together in a crystal can anchor excited nuclei and prevent them from recoiling when they emit their photons. We will not go into the conditions in detail. One condition has to do with the energy of the photon (which determines the speed of recoil). If the energy is too great, the crystal binding-forces cannot hold any of the emitting nuclei, and they tear loose and recoil.

A second condition concerns oscillations that are always set up in a crystal when a nucleus does recoil. According to quantum theory such oscillations occur more easily at higher temperatures. Hence cooling the crystal reduces its ability to accept the recoil energy, and increases the fraction of nuclei that do not recoil.

In Mössbauer's original studies he used the excited nucleus of iridium 191, whose photon has an energy of 129,000 electron volts. In order to anchor even a few per cent of these nuclei he had to cool the crystal to the temperature of liquid air.

After Mössbauer's work became known in this country, experimenters at Harvard University, the Argonne National Laboratory and elsewhere found several other nuclei that exhibit resonance absorption. The most useful is iron 57. The low energy of its photons (14,000 electron volts) and its relatively long half-life (10^{-7} second) give a sharp resonance of considerable intensity at room temperature.

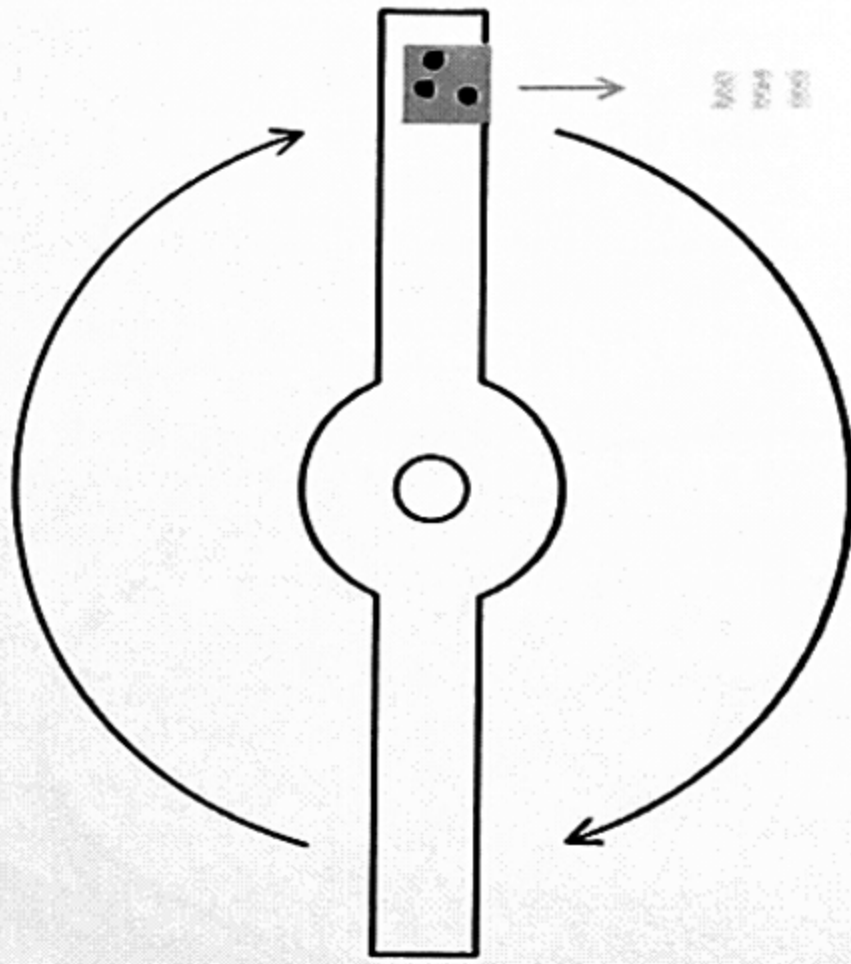
When recoil is eliminated and resonance is established, a slight motion of the source with respect to the absorber produces enough frequency shift to destroy the resonance again. This provides a means of unparalleled sensitivity for observing the Doppler effect. With older instruments the effect could be observed only for rather high velocities, *i.e.*, velocities of the same order as those of the



DOPPLER EFFECT is a change in frequency of waves emitted by a moving source (*dot*) or received by a moving observer. Waves are crowded in the direction of motion and spread in the opposite direction.



100%
100%

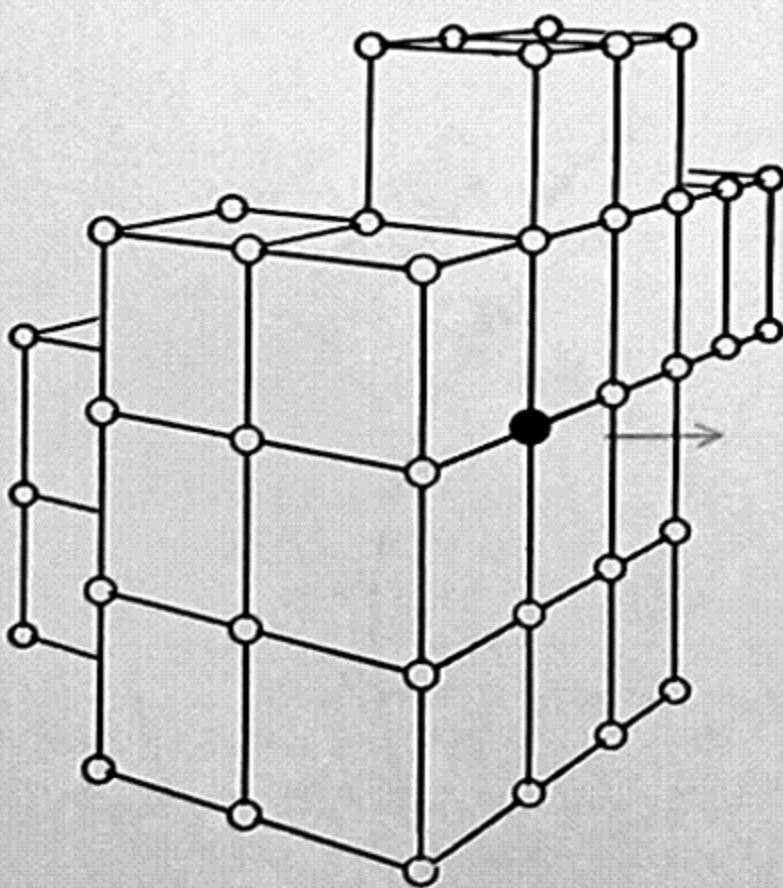
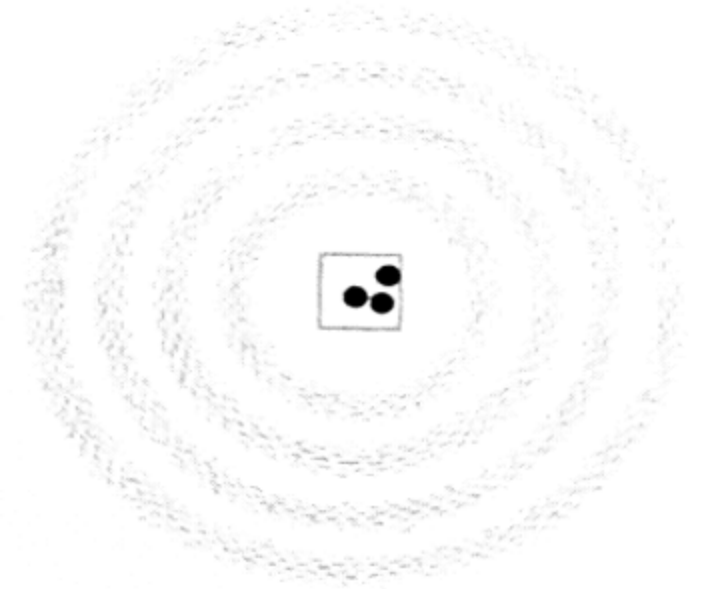


100%
100%
100%

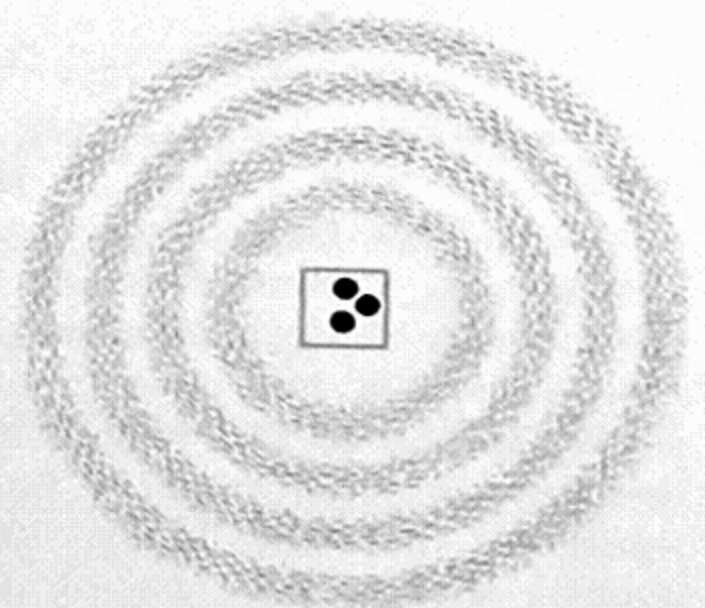
100%
100%
100%

100%
100%
100%

100%
100%
100%



100%
100%
100%



RECOILING NUCLEUS (*dot at top left*) emits photons of less than maximum possible frequency, or energy, which are therefore out of resonance with absorbing nuclei (*dots at top right*). Offset-

ting the recoil velocity by moving the emitter in the opposite direction (*center*) can restore resonance. In the Mössbauer effect (*bottom*) recoil is prevented by anchoring emitting nucleus in a crystal.

waves involved. Fast trains were needed in order to perceive the acoustical Doppler effect; swift astronomical or atomic motions to detect the optical (electromagnetic) Doppler effect. By means of nuclear resonance, Doppler-frequency changes can be observed for ridiculously small velocities. Mössbauer put his source on the rotating turntable of a record player and the resonance was gone! Still more amazing, with iron 57 the resonance disappears at speeds of the order of a couple of inches per minute.

Moreover, the iron resonance-line is by no means the narrowest known. There are isomers whose half-lives are measured in seconds, or even in months, whose line width must be at least a million times sharper. But will it ever be possible to observe the corresponding resonances? To do so it will be necessary to control the relative speed of source and absorber within a hundred millionth of a centimeter per second or less, and this may present difficulties.

In studying the detailed properties of a nuclear resonance the source is moved toward and away from the absorber at various closely controlled speeds, and the intensity transmitted by the absorber is measured. When the transmission for iron 57 is plotted against velocity, the resulting curve shows a large absorption dip at zero relative speed and some smaller dips on either side [see illustration on page 600]. This shape means that the resonance is not actually a single line, but is split into closely spaced lines. The explanation of "hyperfine" splitting, as it is called, is to be found in the magnetic properties of nuclei.

Each nucleus is a small magnet, and the iron-57 nuclei are located in the strong magnetic field of the iron crystal. Quantum theory tells us that the little nuclear magnets can take only certain orientations with respect to the surrounding field; each orientation corresponds to a slightly different energy. Hence every energy level of the nucleus is actually split into several sublevels, corresponding to the various possible orientations in the magnetic field. The transitions between levels, and the emitted photons, have slightly different energies, or frequencies, depending on which sublevels are involved.

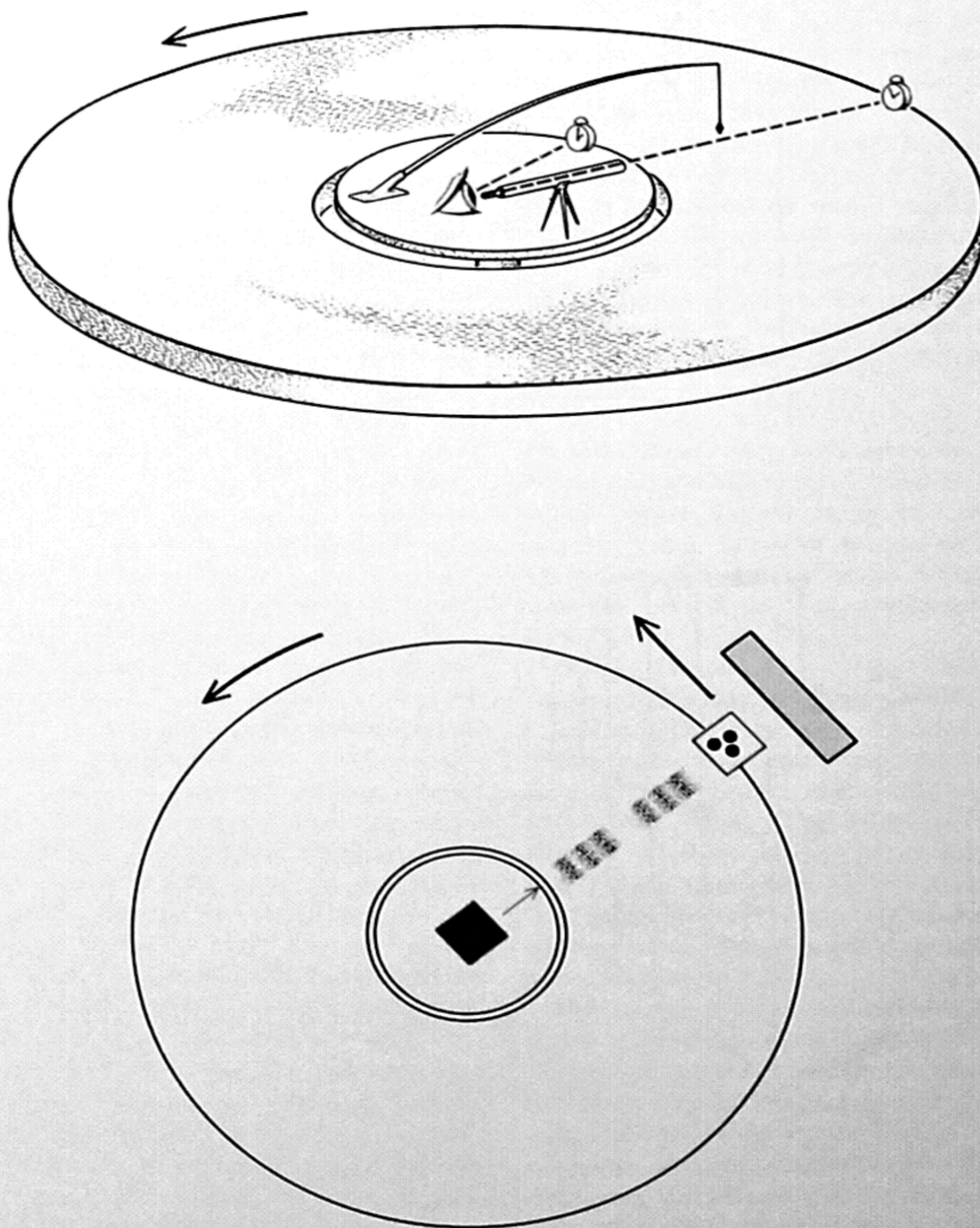
The study of hyperfine splitting has just started. The Argonne group has already obtained valuable information about the magnetization of excited nuclei, which is difficult to measure in other ways, as well as about the magnetic fields existing in certain solids. But

by far the most fascinating applications of nuclear-resonance absorption are those concerning the essence of time. As we remarked at the outset, two experiments suggest themselves: a test of the contraction of clock time due to velocity, as predicted by the special theory of relativity, and of the similar contraction due to gravitational fields, postulated by the general theory of relativity.

According to special relativity the usual concepts of time and simultaneity are valid only for observers who are not in relative motion. If the watch of a stationmaster and that of a train conductor agree perfectly when the train is in the station, they no longer agree when the train moves: each man will see the other's clock as going too slow. The prediction has been verified in the case of

subatomic particles traveling at nearly the speed of light, and the special theory of relativity now rests securely on uncontroversial ground. It is nonetheless of considerable interest to observe the contraction of time in macroscopic bodies moving at comparatively low speeds. Moreover, it should be possible to check one consequence of time contraction that has never been observed with certainty—the transverse Doppler effect.

As we ordinarily think of the Doppler effect, we expect to find a change in frequency in the direction of motion of the radiating source, but not in the direction perpendicular to the motion [see bottom illustration on page 596]. However, relativity shows that there must be a small decrease in frequency in the transverse direction, because to a stationary observer the vibrations of the moving source

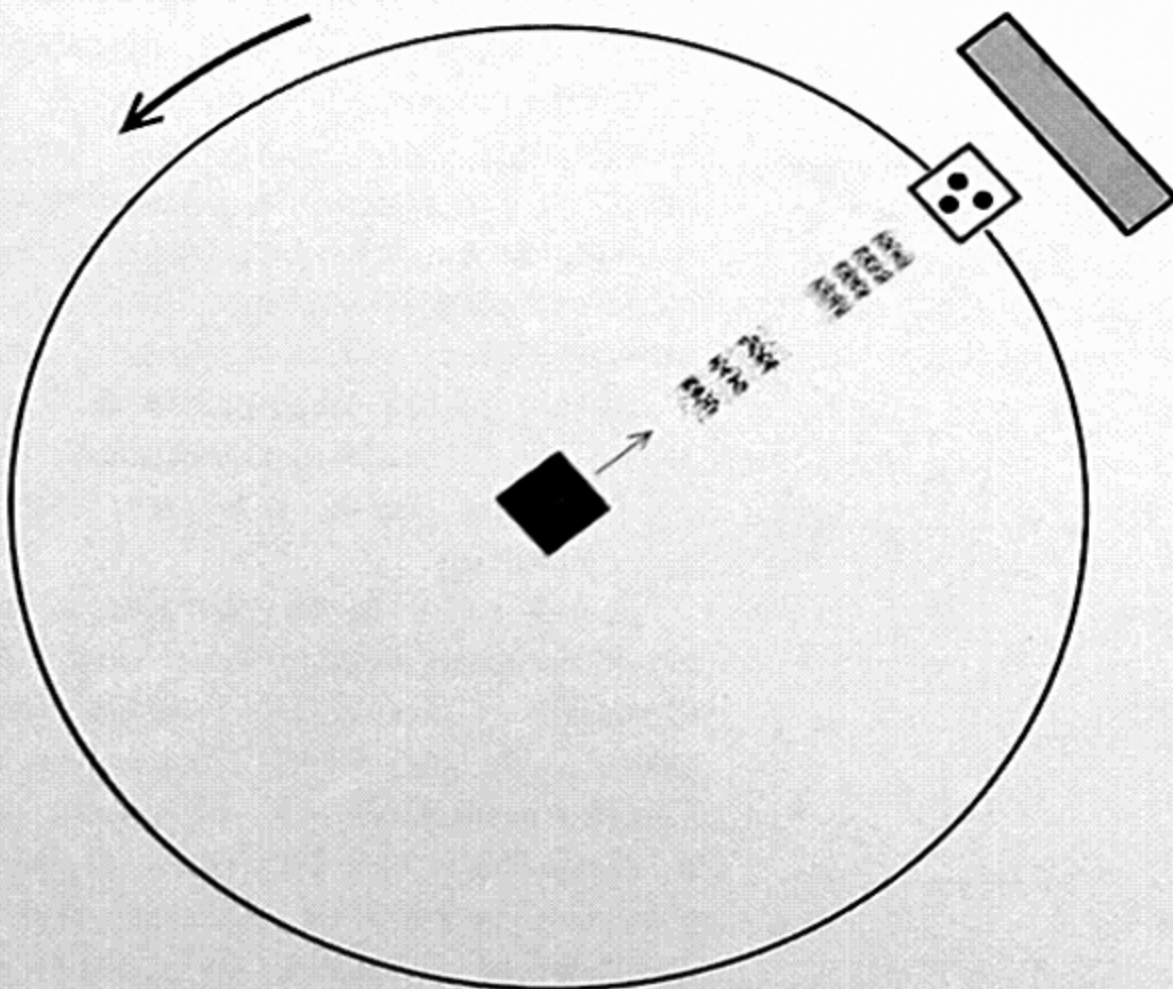
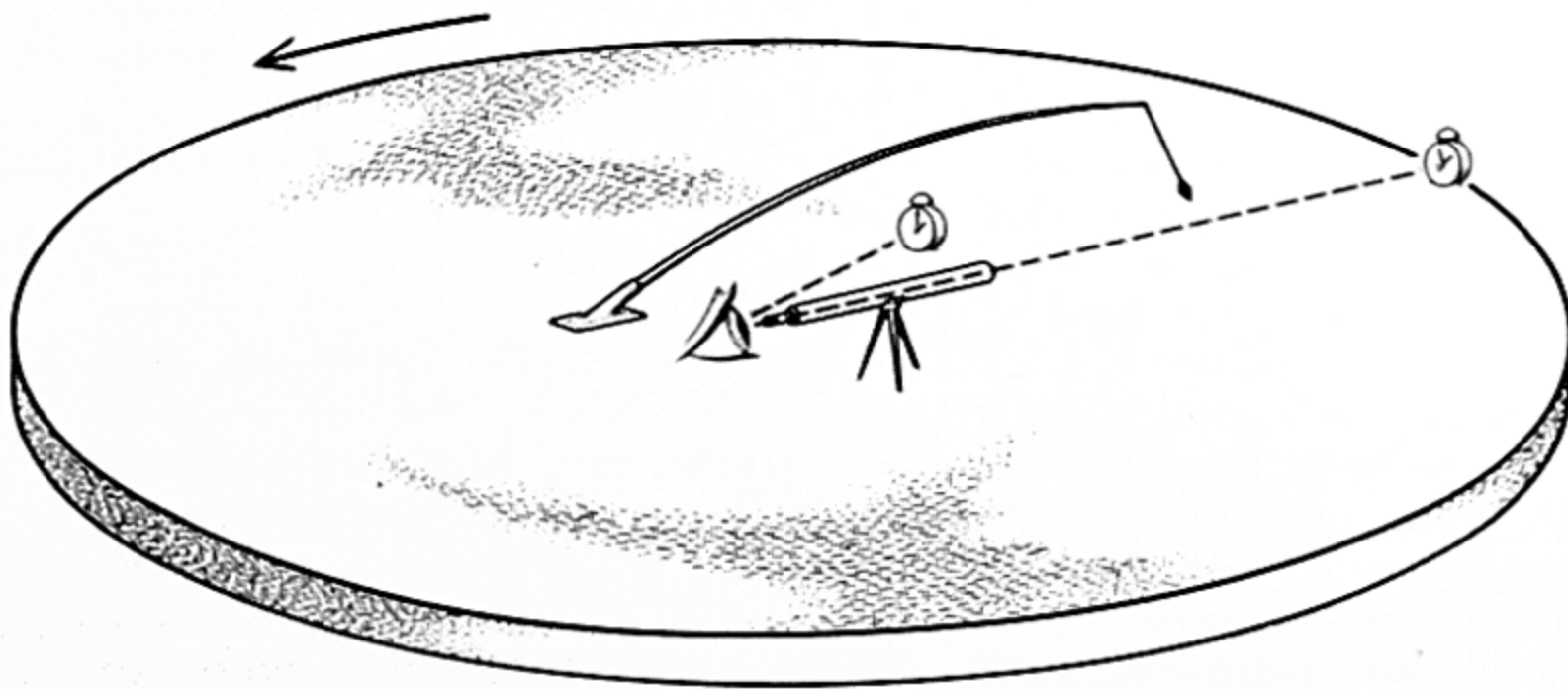


RELATIVISTIC EFFECTS are illustrated schematically in these diagrams. Drawings at top represent imaginary experiments comparing clocks that are in relative motion (left and

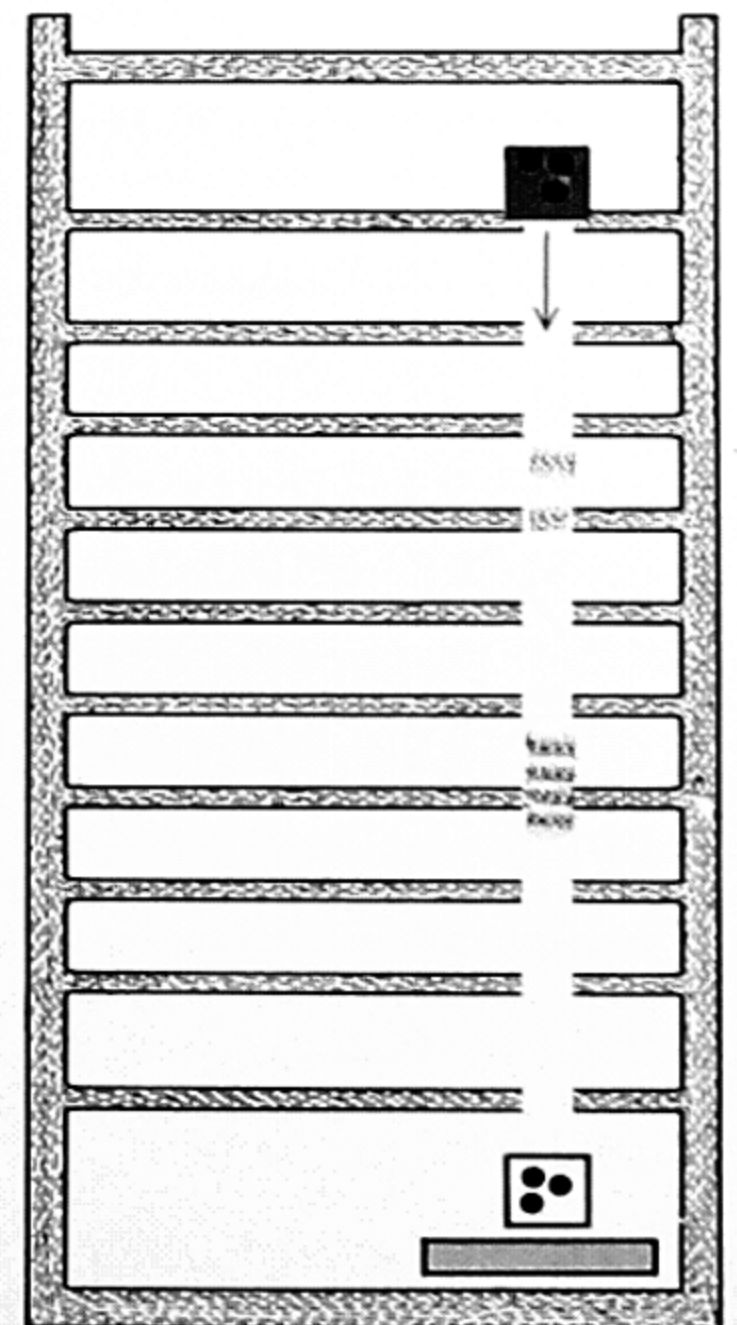
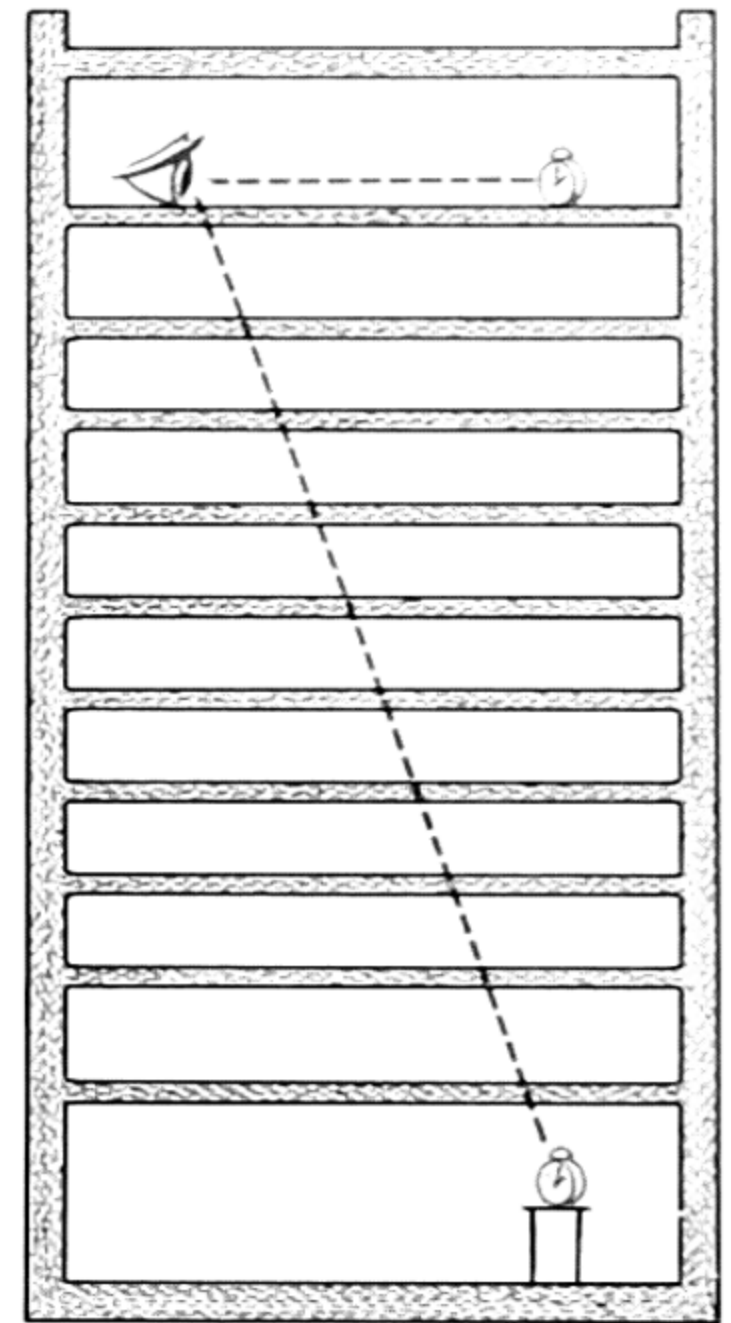
will seem too slow. An experiment to measure the transverse Doppler effect has been conducted in England by H. J. Hay, J. P. Schiffer, T. E. Cranshaw and P. A. Egelstaff. They placed a source of iron-57 nuclei at the center of a rotating disk, put an absorber on the rim and a detector beyond the disk [see illustrations on these two pages]. When the disk was rotated at 500 revolutions per second, the counting rate was 4 per cent higher than when it was stationary, indicating that the absorber was out of resonance with the source because of the transverse motion. This is what one expects from time contraction.

Let us look a little more closely at the interpretation of this experiment. The iron nuclei at the center and at the rim of the disk are clocks, or at least pendulums, and we are trying to compare their

frequencies of oscillation. To visualize the situation more vividly let us transform the disk into a merry-go-round and put a human observer at the center. With him he has a clock and a fishing rod. On the edge of the merry-go-round is a second clock, identical with the central one. Now suppose that the observer, together with his equipment, is on a platform that does not turn with the merry-go-round. To keep the distant clock in sight he must himself turn around. Thus he knows that the clock is moving, and when he sees it running slower than his local clock, he attributes the difference to the time contraction of special relativity. His fishing line, hanging straight down in the space between his stationary platform and the merry-go-round [see top illustration on opposite page] indicates that there is no horizontal force



center) or at different gravitational potential energies (right). Lower drawings show how nuclear resonance-absorption measure-



ments, as described in the text, make such experiments possible. Turntables test special, and tower general, theory of relativity.

acting between the two clocks. We shall see the significance of this observation in a moment.

If we replaced the ordinary clocks with "clocks" of iron-57 nuclei, the observer could not actually perceive that one nuclear clock vibrates more slowly than the other. What he would perceive is that the frequency of the photons sent out by the central clock is different from the frequency of the distant clock, or absorber. Since he believes the photon to have the "right" frequency, he says that the absorber frequency is too slow.

Now imagine a slightly different situation. The observer and his clock (or source) are no longer isolated from the merry-go-round, but turn with it. Would we expect a transverse Doppler effect here? The answer is yes, because the linear speed connected with the observer's rotation at the center of the disk is much less than the speed of the absorber on the rim. But in this case an entirely different interpretation is possible.

To return to the man and his fishing rod, he can now stand still and keep the distant clock in sight; it appears stationary to him. But when he looks at his fishing line, he finds it hanging outward, toward the rim of the merry-go-round. Thus he perceives a horizontal force that pushes things from the center to the outside. It is, of course, the inertial (centrifugal) force arising from the fact that the merry-go-round does not move at

constant speed in a straight line; in other words, the motion is accelerated. He now attributes the disagreement between clocks (actually, the lack of resonance) to the presence of the force.

He reasons as follows: The photons emitted by the source, in traveling toward the absorber at the rim, must be acted on by the outward force that is revealed by the fishing line. Since the photons are already moving at the maximum possible velocity—the velocity of light—the force cannot speed them up, but it does increase their energy and therefore their frequency. Thus they arrive at the absorber with too high a frequency, and the resonance is destroyed.

Note that in the first case the observer believes that the frequency of the photons is constant and that the frequency of the absorber decreases. In the second case he thinks that the absorber remains at constant frequency while the photon frequency increases. Experimentally there is no way to decide which has happened. They both result in the same detuning from resonance. Furthermore, computation shows that the difference in frequency is the same either way.

Thus far we have given two equivalent explanations of an effect of special relativity. But now we pass to the next question: Does gravity act on photons in the same way that centrifugal force seems to? The answer is not obvious. For example, electric and magnetic fields do not act on photons at all. How can we

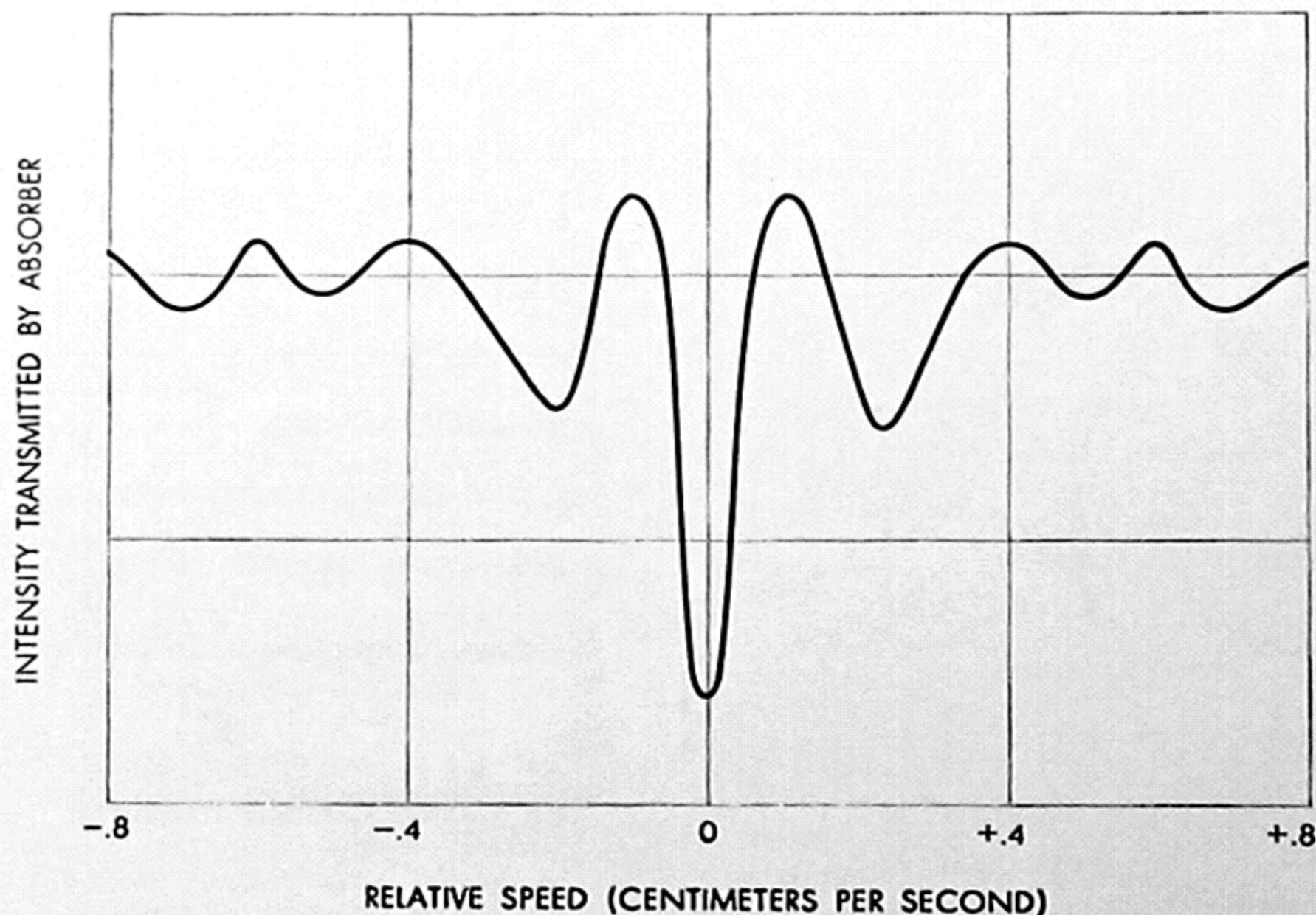
decide *a priori* whether gravity will? Einstein answered this question. The fundamental assumption of the theory of general relativity is that gravity acts in exactly the same way as inertial forces that appear in accelerating systems. The assumption is known as the principle of equivalence. It predicts that there will be a lack of synchronization in clocks that is caused by gravitational fields, or, to put it another way, that photons being pushed or pulled by gravitational forces will change in frequency.

Modern physicists are inclined to believe in the validity of general relativity for esthetic reasons, because it is mathematically so elegant and philosophically so satisfying. They use the theory in all speculations on cosmology, including the questions of the curvature of the universe, its size, its beginning, its expansion and its evolution. But although certain experiments seem to check some consequences of the principle of equivalence, according to the experts there is no verification that stands a strict critical analysis.

Presently it seems as though nuclear resonance should provide the long-awaited crucial test. The experiment is very simple in principle. A source of iron 57 is allowed to emit photons vertically, say in the downward direction. Do these photons gain energy, and thus frequency, while traveling down, because of the gravitational attraction of the earth? Is there a detuning, a clock contraction, with an absorber a few stories lower? The expected effect, if general relativity is correct, is small—of the order of one hundredth of the line width. But it should be possible to observe it with available instrumentation.

Schiffer, Cranshaw and A. B. Whitehead, who have performed the experiment at the British Atomic Energy Establishment in Harwell, report that their initial results show an effect about 96 per cent as great as the theory predicts, with an error of 45 per cent. R. V. Pound and Glen A. Rebka, performing the test at Harvard, have not yet announced their findings.

In any case the matter should be definitely settled before very long. And so Chronos is overthrown along with the other gods. Or is he? The writer, an obstinate humanist and classicist, prefers to think that the triumph of physics is just the victory of Athena, the virgin goddess of intellect, over most of her Olympian companions. If only we do not throw away other human values, if we can save Aphrodite, the fertile goddess of beauty and love, all should be well.



RADIATION TRANSMITTED by an absorber moving with respect to a source is plotted against the relative velocity of the bodies toward one another (positive) or away (negative). The multiple dips in the curve represent hyperfine splitting due to nuclear magnetism.

The Author

SERGIO DeBENEDETTI is professor of physics at the Carnegie Institute of Technology. He was born in Florence in 1912, and studied in the Laboratory of Arcetri, near the hill where Galileo died. He took his Ph.D. at Florence, where he was associated with Bruno Rossi, the cosmic ray expert. In 1938 he left Italy for the Curie Laboratory in Paris, and there developed an interest in positrons. DeBenedetti has been in the U.S. since 1940; in 1946-1948 he was principal physicist at the Clinton Laboratories in Oak Ridge. His interest in elementary particles has led him to the special, and enjoyable, occupation of making atoms with new particles. As for hobbies, he says: "The thing I really like is to see the world, and with the excuse of physics and of cosmic rays I have touched all continents but Asia and Australia, a situation which I intend to remedy before too long. Of course I am interested in world affairs. I also like to draw and paint."

Bibliography

GRAVITATIONAL RED-SHIFT IN NUCLEAR RESONANCE. R. V. Pound and G. A. Rebka in *Physical Review Letters*, Vol. 3, pages 439-441; November 1, 1959.

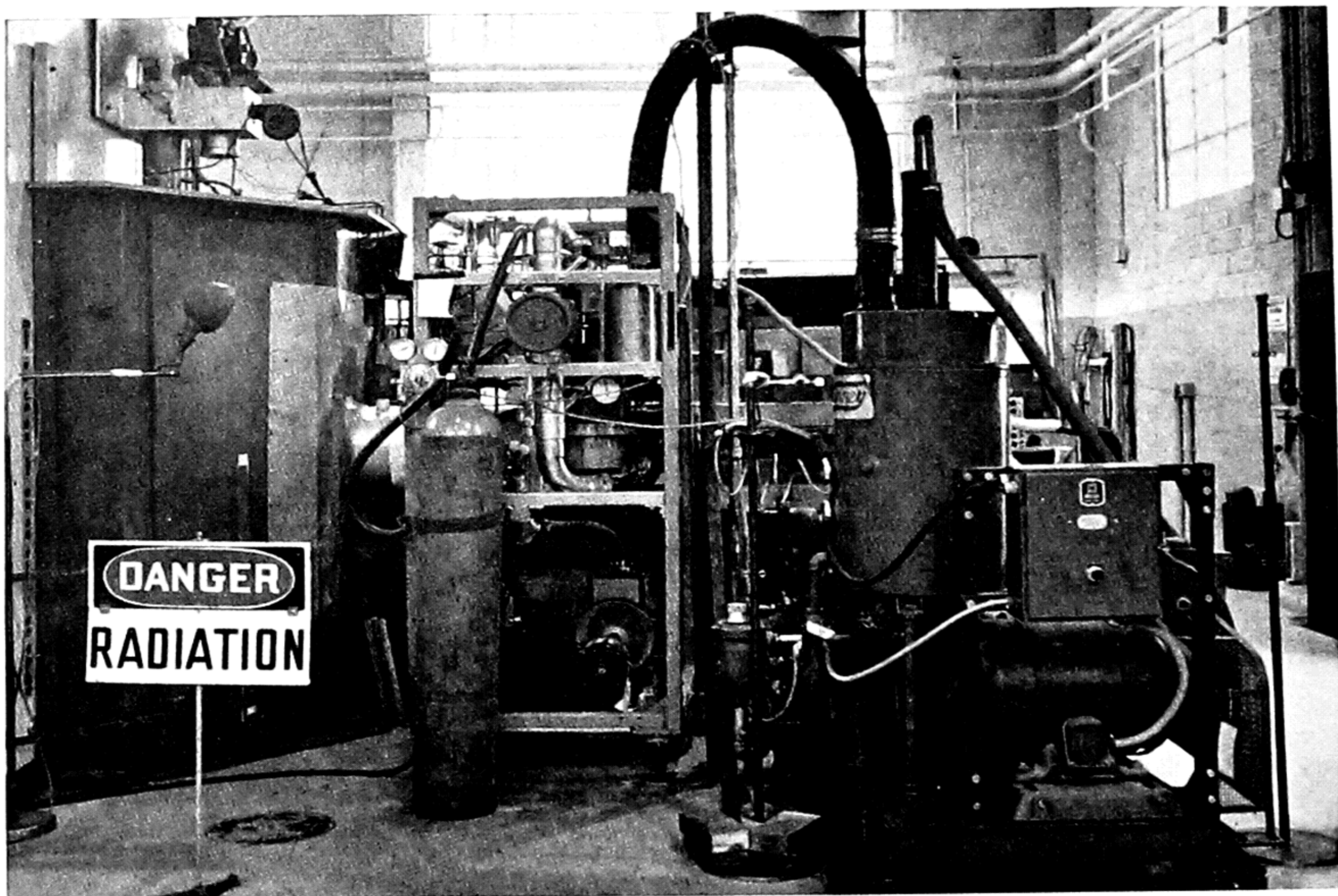
KERNRESONANZFLUORESZENZ VON GAMMASTRAHLUNG IM Ir^{191} . Rudolf L. Mössbauer in *Zeitschrift für Physik*, Band 151, Heft 2, pages 124-143; 1958.

RESONANT ABSORPTION OF THE 14.4-KEV γ -RAY FROM 0.10- μ SEC Fe^{57} . R. V. Pound and G. A. Rebka in *Physical Review Letters*, Vol. 3, pages 554-556; December 15, 1959.

RESONANT SCATTERING OF γ -RAYS. K. G. Malmfors in *Beta- and Gamma-Ray Spectroscopy*, edited by Kai Siegbahn, pages 521-530. North-Holland Publishing Company, 1955.

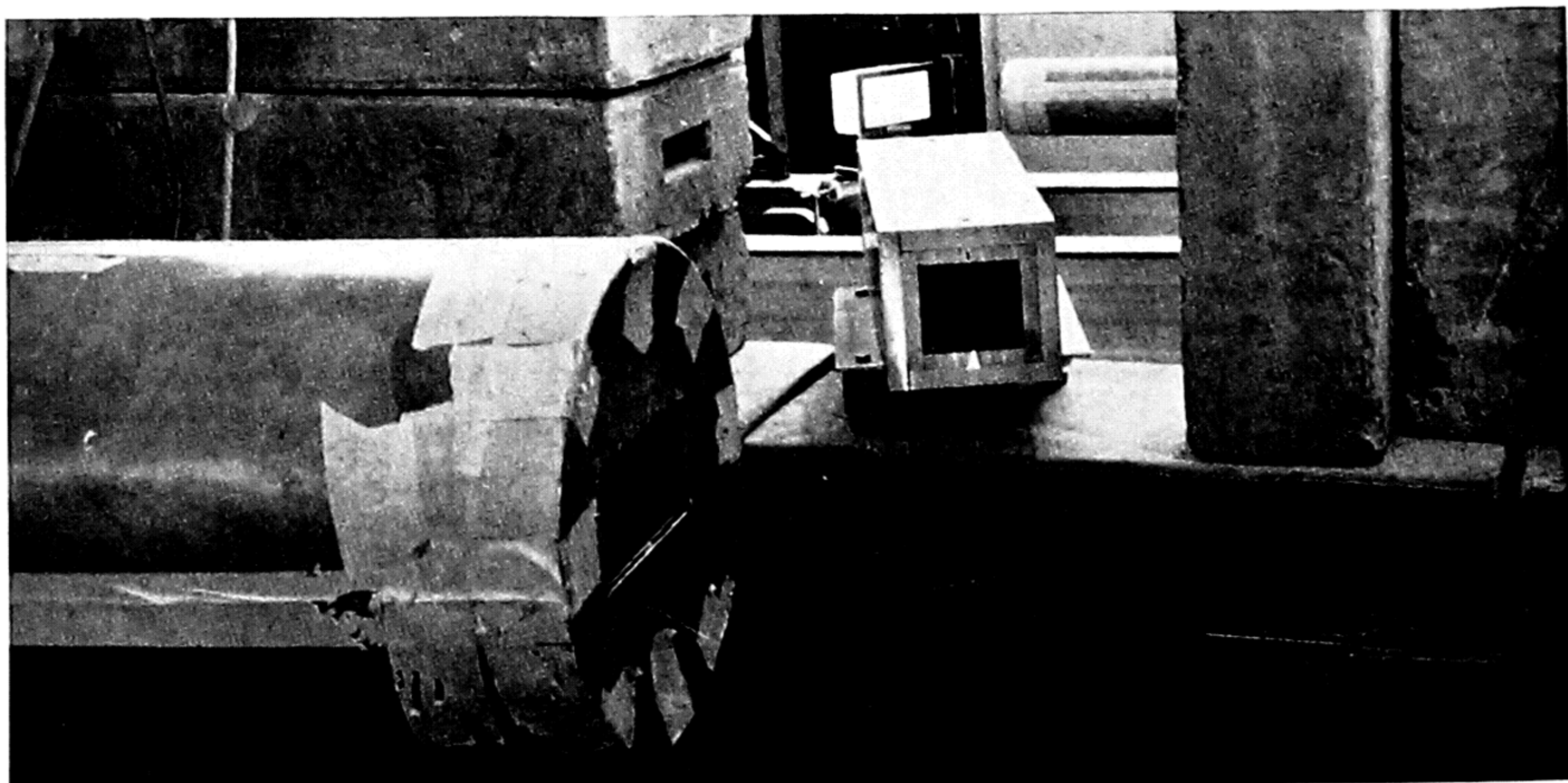
THE THEORY OF RELATIVITY. C. Møller. Oxford University Press, 1952.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.



SCATTERING OF NEUTRONS by superfluid helium is measured in this apparatus at the Los Alamos Scientific Laboratory by John

L. Yarnell. Neutron source is Omega West reactor at far left. Liquid helium is in small cylinder behind pipe-elbow at left center.



TARGET AREA is shown with helium target removed. Neutron beam enters through opening at left. Target (normally located at

center) scatters neutrons toward collimator (center). Neutrons are detected by scintillation crystal (top center) and counters.

SUPERFLUIDITY AND "QUASI-PARTICLES"

by F. Reif

The curious properties of liquid helium are explained by regarding it as a gas of hypothetical particles in a similarly hypothetical background fluid. Recent experiments strikingly confirm the model.

It is a cliché to say that theoretical physics has become hopelessly abstract and does not provide "models" that can be intuitively understood. That is why the story of liquid helium seems to me to have a special appeal. It provides a better glimpse than most into the way in which a few simple, abstract ideas can provide the key to a concrete but baffling problem.

Helium is a remarkable substance. Its properties were described in *SCIENTIFIC AMERICAN* not long ago by Eugene M. Lifshitz ["Superfluidity"; Offprint 224]. Here I shall mention only a few: The temperature at which helium liquefies—4.2 degrees absolute (−269 degrees centigrade) at atmospheric pressure—is lower than that for any other gas. Unlike all other substances, helium does not freeze; it remains liquid down to absolute zero.

At temperatures not too far below 4.2 degrees absolute, helium behaves like an ordinary liquid. For example, when it is boiling, bubbles are formed in its interior. But below 2.18 degrees—the so-called lambda point—the bubbling suddenly ceases and the liquid becomes perfectly still. The transition is marked by abrupt changes in other properties of the substance; for example, in its heat capacity. Above the lambda point the liquid is called helium I; below it, helium II.

Helium II has an uncanny ability to pass through extremely small openings: it flows easily through a slit less than a hundred thousandth of an inch wide, which would be virtually impassable to water. Accordingly it is said to be "superfluid." The speed with which an ordinary liquid moves through a narrow tube or slit is inversely proportional to its viscosity; water flows faster than molasses. On this basis helium II is incomparably less viscous than any oth-

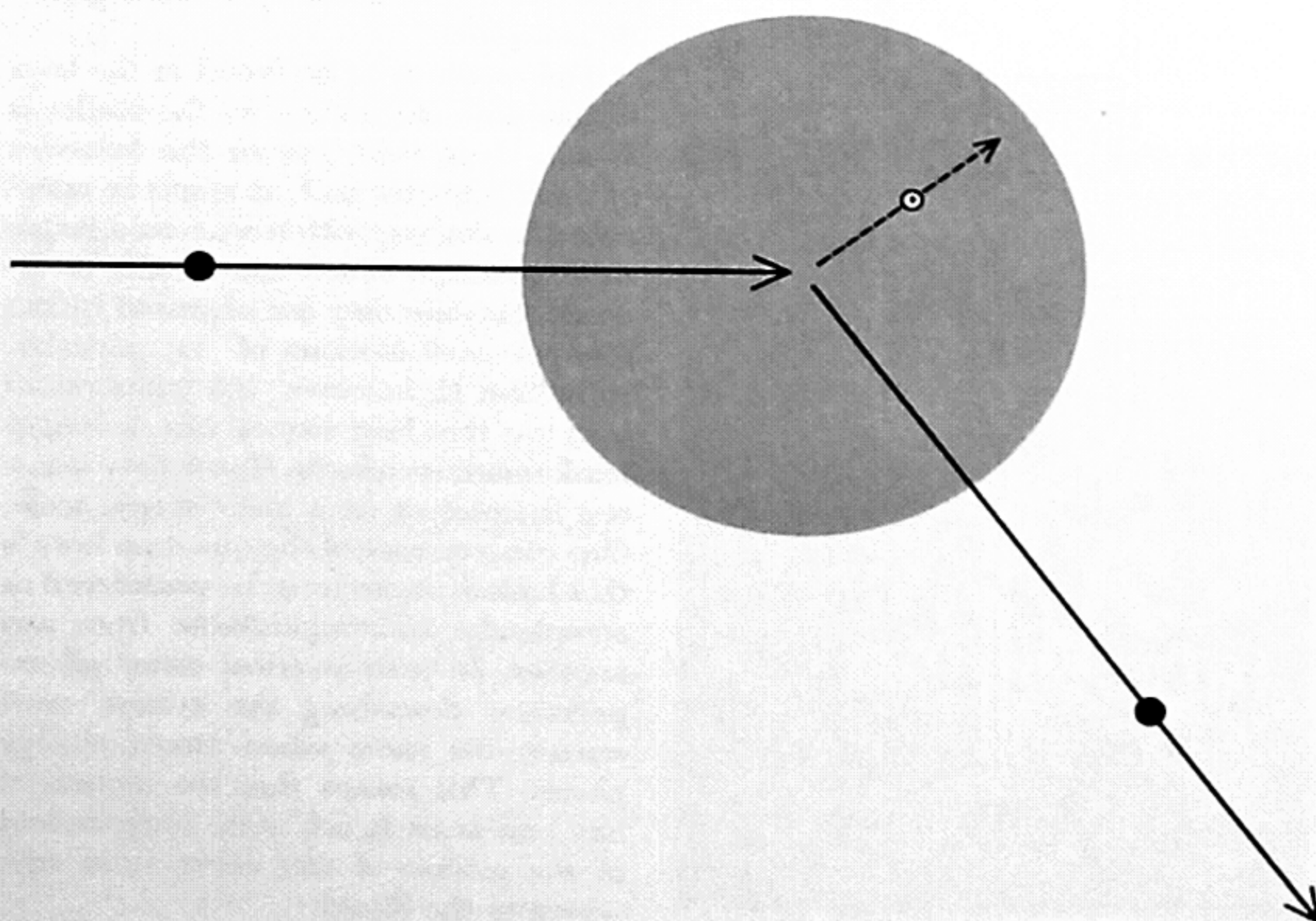
er liquid; in fact, its viscosity is less than a ten thousandth that of hydrogen gas.

Another common method of determining viscosity is to enclose a fluid in a narrow gap between two concentric cylinders. When the outer cylinder is slowly rotated, the fluid transmits a force to the inner cylinder, tending to turn it in the same direction. This force is by definition a measure of the viscosity of the fluid. Measured in this way, the viscosity of helium II turns out to be appreciable; at some temperatures it is greater than that of "normal" helium I!

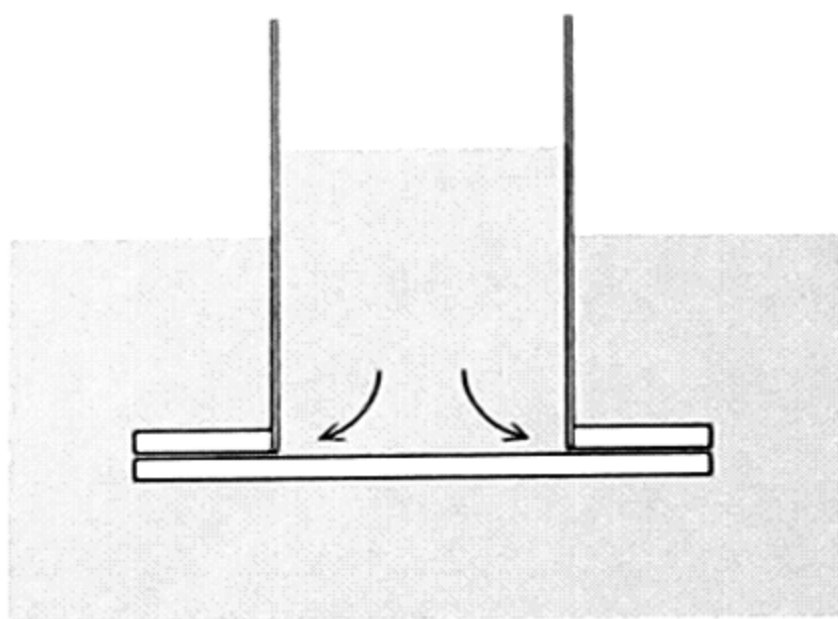
For the present I shall mention just one more feature of this strange liquid: the fact that local differences in temperature can produce pronounced mechanical effects. When a vessel with a

very narrow neck is partially immersed, with its open end down, in a bath of helium II, this liquid, like any other, flows into the vessel until the level is the same inside and outside. But if the liquid in the vessel is heated, the flow starts again, and the level inside the container rises higher than that on the outside [see illustration at bottom left on next page]. If such a vessel also has an opening at the top, helium II can be made to spurt out the top like a fountain. This phenomenon is therefore known as the "fountain effect."

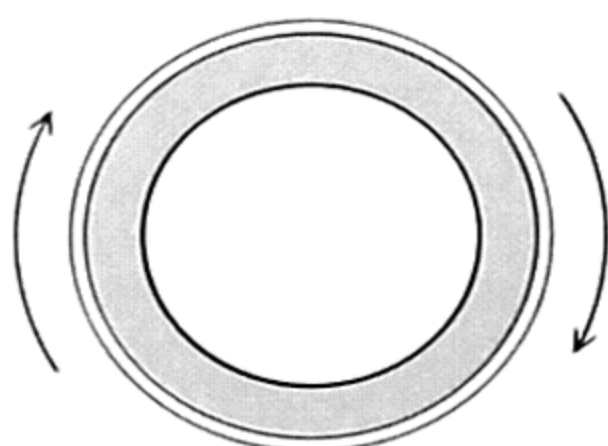
The list is not exhaustive, but it gives a fair idea of what the theoretical physicist was called on to explain. Before turning to theory, let us take a moment to consider what "explaining"



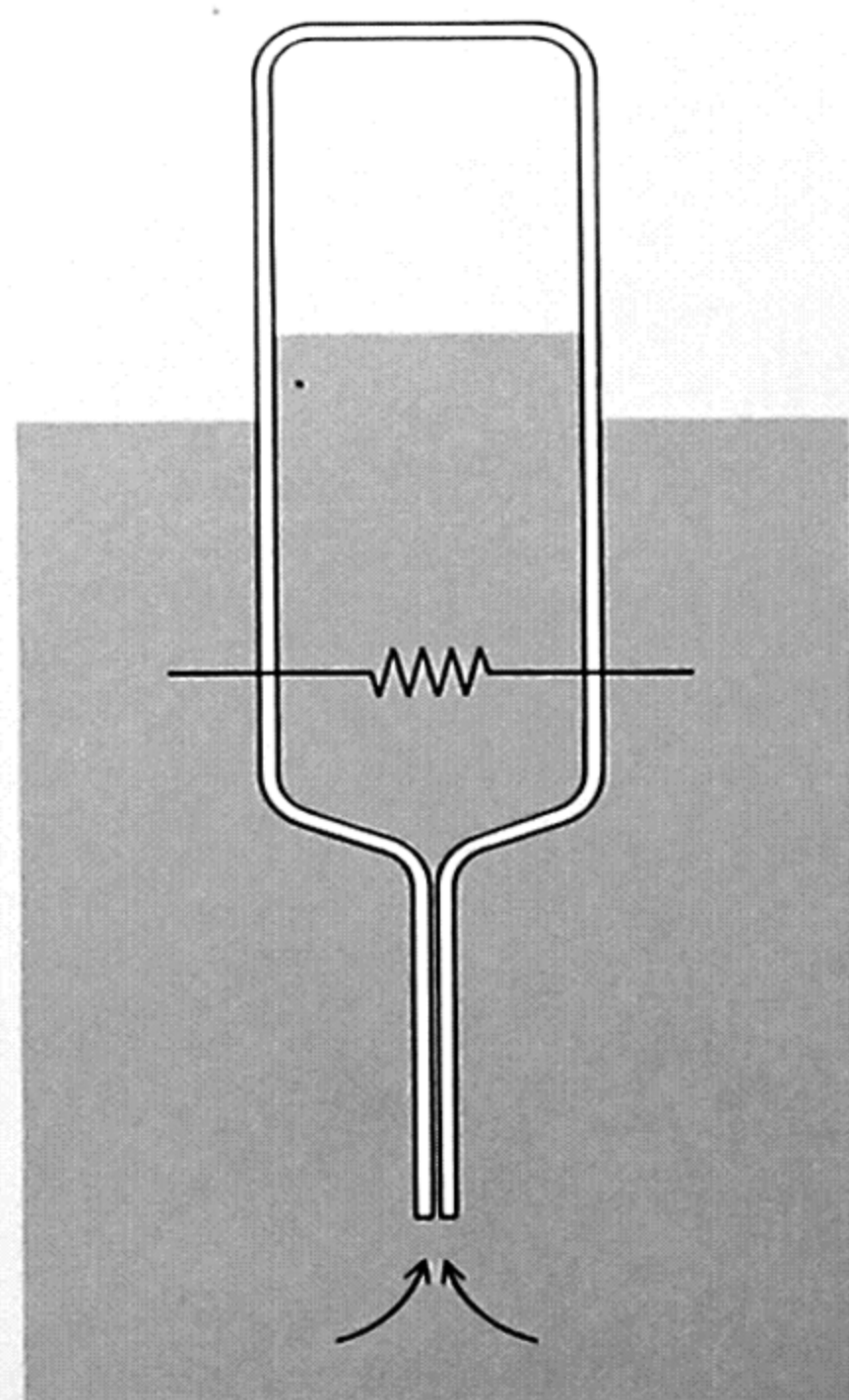
QUASI-PARTICLE (top right) is created when superfluid helium (color) scatters a neutron (black ball). Quasi-particle acquires the energy and momentum lost by neutron.



SUPERFLUIDITY of helium II (*color*) makes it possible for the liquid to flow freely through a very narrow channel formed by two polished glass plates. Such a channel is virtually impervious to the flow of water.



VISCOMETER consists of two concentric cylinders. When outer cylinder is rotated, the fluid (*color*) exerts a twisting force on the inner cylinder. The magnitude of the force depends on the viscosity of the fluid.



FOUNTAIN EFFECT causes level of helium II in vessel to rise when liquid in vessel is heated. Heater is shown at center.

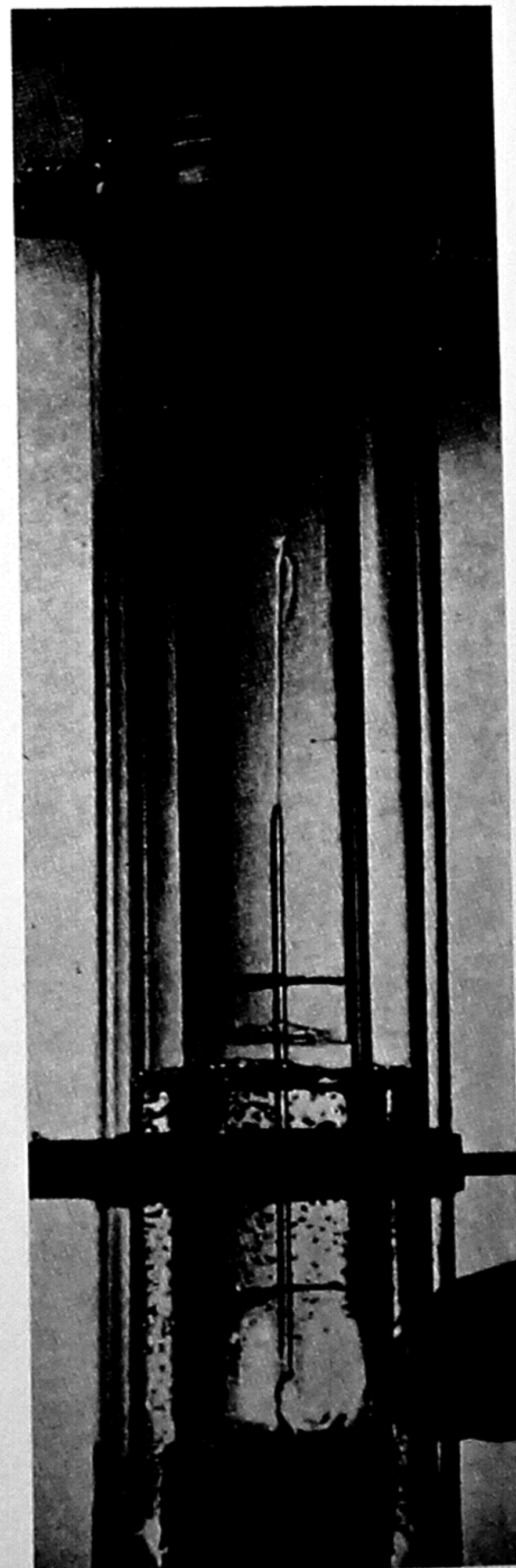
means. The physicist seeks to account for the gross properties of matter in terms of the behavior of its constituent particles. The program has had its greatest success in dealing with dilute gases, where the individual molecules are so far apart that they spend most of their time moving as isolated bodies. When they do meet, they come together two at a time; the chance of a simultaneous collision of more than two molecules is negligible. Thus the theory of gases can be reduced to a tractable "two-body" problem. Liquids, even the ordinary well-behaved kind, pose a much more difficult problem. Their molecules are so densely packed that each one is always under the direct influence of about a dozen neighbors. Considering that even the celebrated three-body problem of mechanics still presents formidable difficulties, it is little wonder that the theory of liquids is far from satisfactory. Nevertheless, by making drastic simplifications and focusing attention on the motion of at most a few molecules at a time, it is possible to understand the behavior of ordinary liquids at least semiquantitatively.

At first glance liquid helium would seem to be particularly easy to deal with in this way. Its molecules are identical single atoms, and the force they exert on one another is both simple and weak. But we cannot isolate one atom of helium II, even in theory. Not only must we take its immediate neighbors into account, we must consider all the trillions of trillions of atoms in the entire sample simultaneously, as if they made up a single gigantic molecule.

The reason is to be found in the laws of quantum mechanics. As the reader is aware, these laws govern the behavior of small particles such as atoms or molecules. In dealing with matter on a larger scale quantum effects can usually be ignored, because they are obscured by the random heat motions of the particles. In helium II, however, the temperature is so low that heat motion can no longer mask quantum effects. Hence they manifest themselves on a macroscopic scale. One consequence of the quantum laws is that helium atoms must be considered as completely indistinguishable from one another. In mathematical terms all expressions describing the system must remain the same when atoms change places. This means that the motion of any one atom is not quite independent of the motion of any other atom anywhere in the liquid.

But if 12 molecules are too many to consider at a time, what can be done with 10^{23} particles? It turns out that,

near absolute zero, they can be dealt with in a simple and useful way. If we consider the whole liquid as we would a molecule, then at absolute zero it would be in its "ground state" of lowest possible energy. (Quantum mechanics tells us that this energy is not zero. That



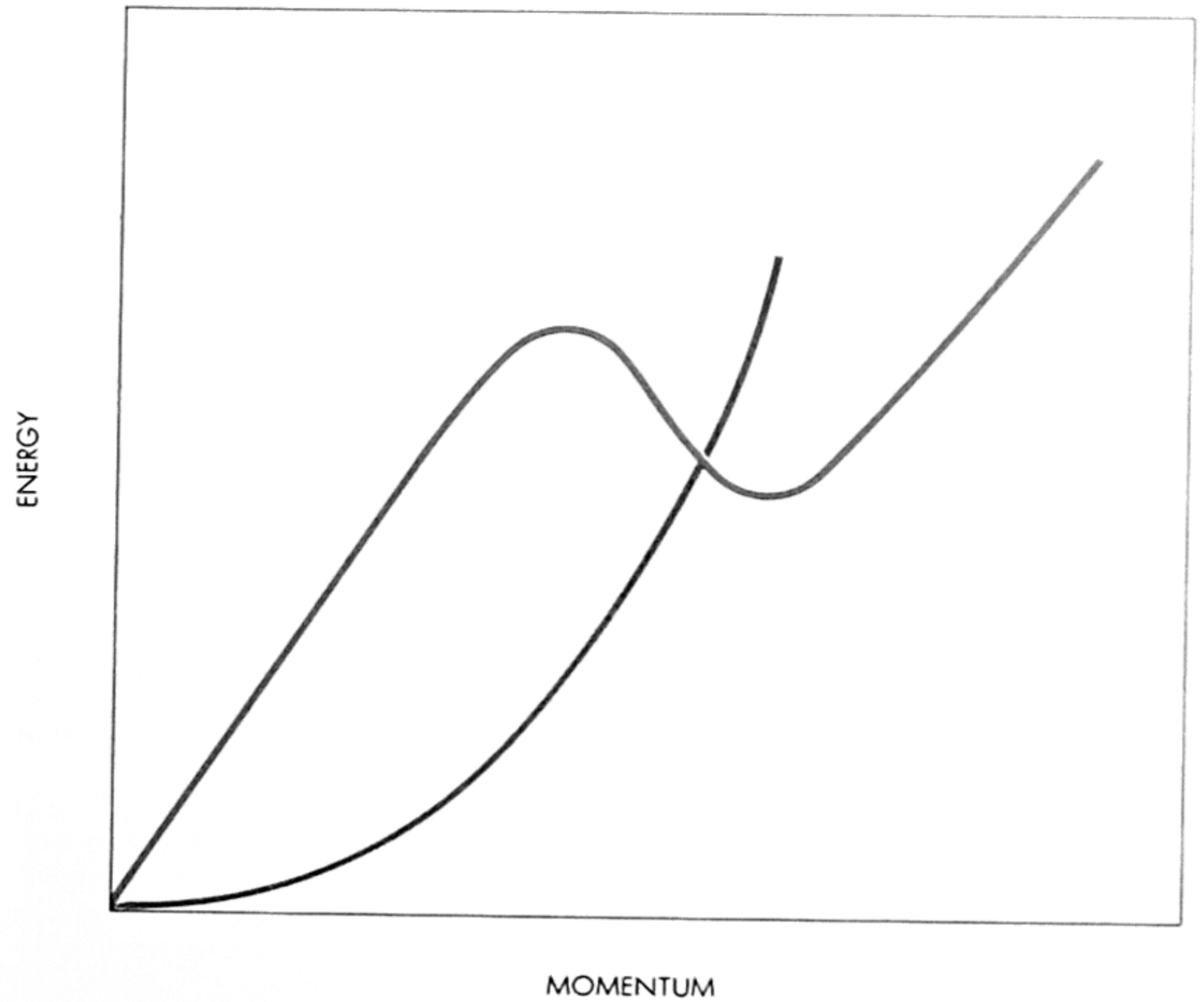
FOUNTAIN of helium II rises from the capillary tube within the double Dewar flask at center. Fountain is caused by the excitation of liquid helium by a light source.

is why the liquid does not solidify.) At higher temperatures the liquid would be found in one of its "excited states" of somewhat higher energy. An excited state implies some sort of motion or vibration in the fluid. At temperatures not too far above absolute zero, the possible excited states are few in number and have low energy. Therefore only comparatively simple types of motion need be considered.

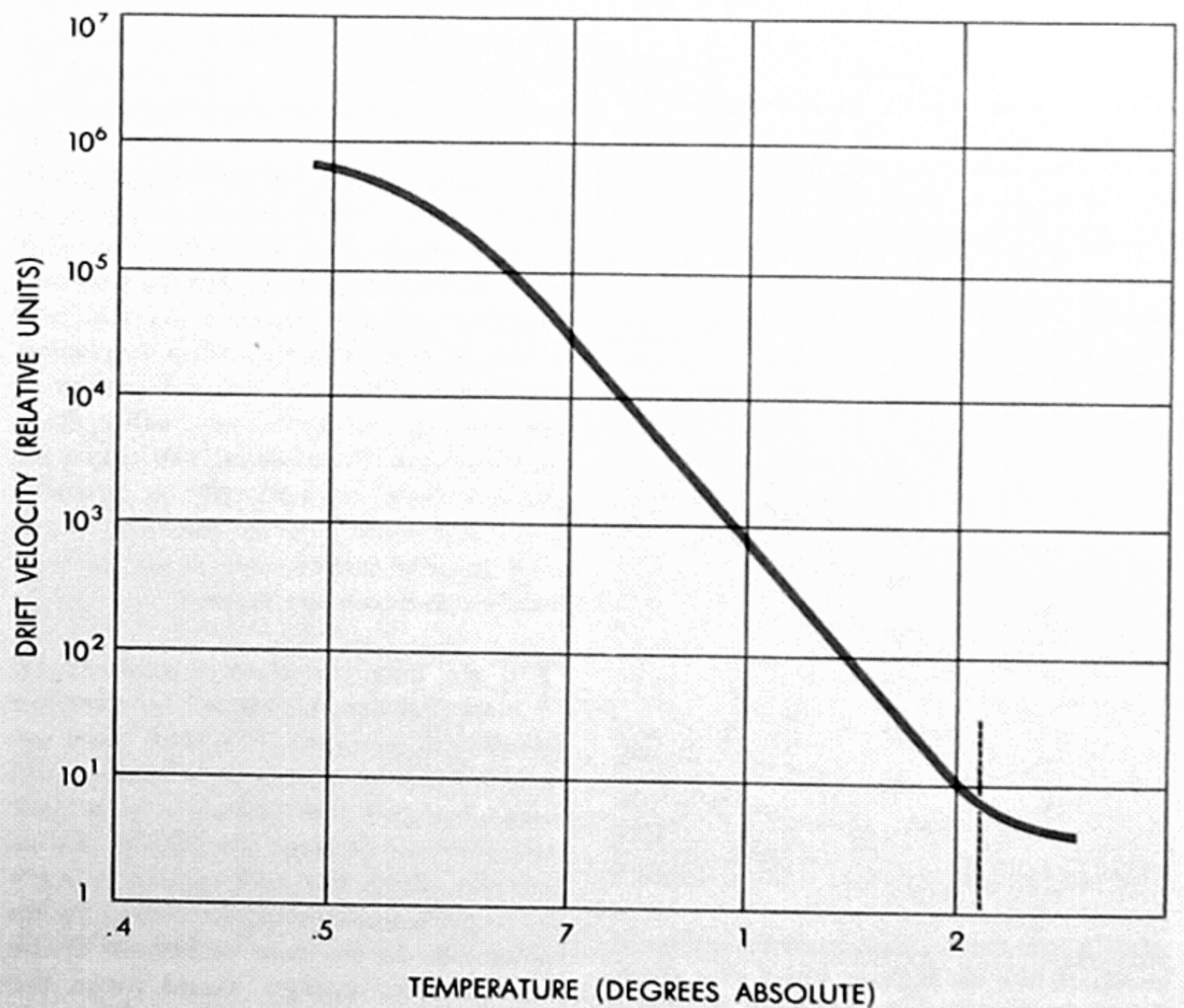
This method of attack was initiated by the Russian theoretical physicist L. D. Landau. He began by trying to find the simplest modes of motion of the liquid as a whole. For example, one such type of mode is a sound wave passing through the liquid. Each sound wave or other mode of motion carries with it a certain amount of energy and momentum. Landau's analysis showed that the energy and momentum of any low-lying excited state must be simple combinations of the values associated with the separate possible modes. Thus the formula for energy is just: $E = E_0 + n_1 e_1 + n_2 e_2 + n_3 e_3 \dots$ E_0 is the energy of the ground state (at absolute zero); the e 's represent the energies of the successive modes of motion. The laws of quantum mechanics restrict the n 's to whole numbers: 0, 1, 2, 3 and so on. Exactly the same relation holds for momentum (denoted by P): $P = P_0 + n_1 p_1 + n_2 p_2 + n_3 p_3 \dots$

A glance at these simple formulas suggests a concrete and useful interpretation of the abstract argument. Note that the expressions for the total energy and momentum of the liquid are the same as if it consisted of a "background fluid" of energy E_0 and momentum P_0 , with a number of individual particles immersed in it. There would be n_1 particles of energy e_1 and momentum p_1 , n_2 particles of energy e_2 and momentum p_2 , and so on. Of course there really are no such particles; to underscore the fact, let us call them "quasi-particles." But all the pertinent equations can be interpreted as though they were actually there. What makes this view so fruitful is that, at low temperature, there are not too many modes of motion to deal with, and hence not too many quasi-particles. They can therefore be considered as making up a dilute gas. Thus the hopeless problem of dealing with the interdependent motions of almost countless numbers of helium atoms is overcome, and can now be treated by the well-understood theories of gases applied to quasi-particles.

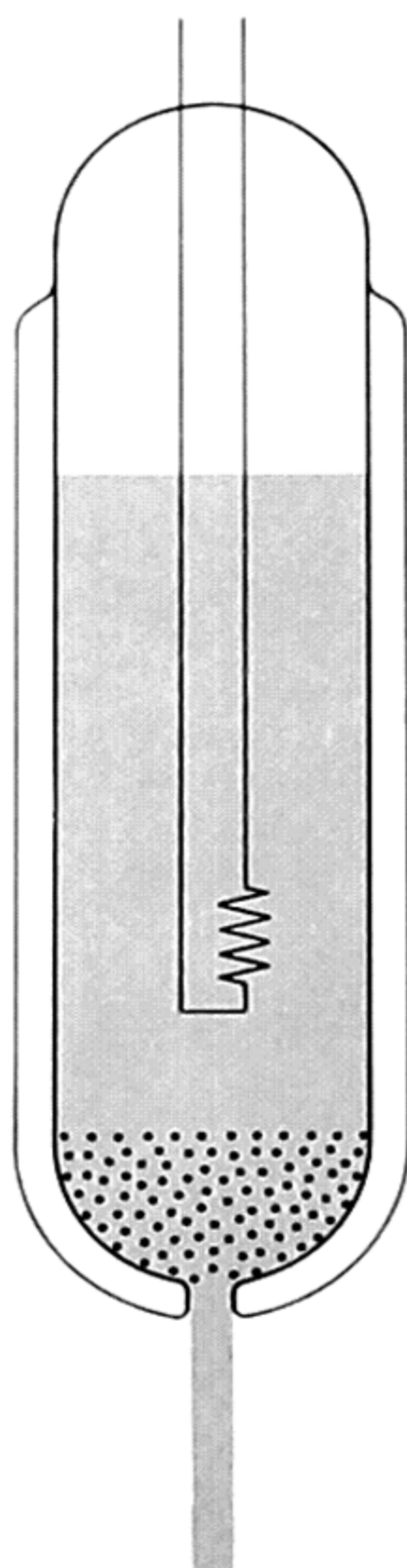
It should be emphasized that a quasi-particle is a purely theoretical construct,



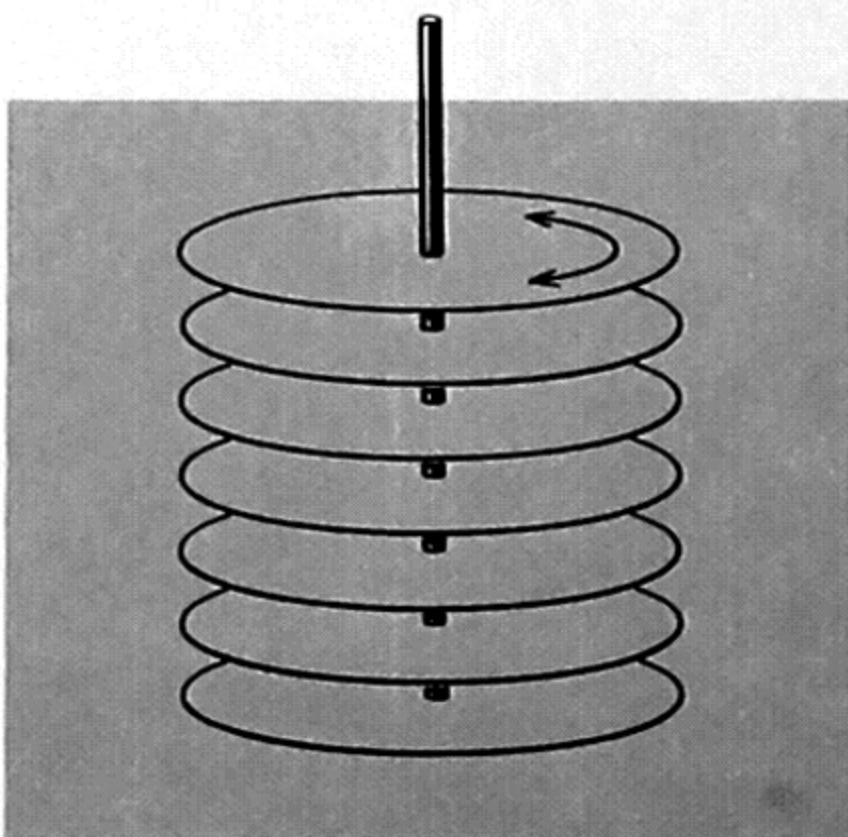
RATIO OF ENERGY TO MOMENTUM for real particles (*gray curve*) can have any value greater than zero. In contrast, the slope of the curve for quasi-particles (*color*) indicates that the ratio of energy to momentum can never be less than a certain minimum value.



DRIFT VELOCITY of a charged particle (ion) moving through helium II increases as the temperature decreases. Above lambda point (*broken line*) velocity is almost constant.



FILTER of finely packed emery powder (dots) separates quasi-particles from helium II. As helium flows out, concentration of quasi-particles increases, raising temperature of fluid in vessel. Electrical-resistance thermometer at center measures temperature.



STACK OF DISKS immersed in helium II oscillates like the balance wheel of a watch. As temperature rises, numbers of quasi-particles adhere to the stack, increasing its mass and hence its period of oscillation.

having to do not with an individual helium atom, but with motion of the liquid as a whole. Nevertheless, the behavior of a quasi-particle gas is remarkably like that of a real gas, with two important differences. The first of these concerns the relation between the energy and momentum of a quasi-particle. This relation reflects the properties of the modes of motion of the whole liquid and is quite unlike the corresponding relation for a real particle. The second difference concerns the number of particles in a given sample of material. In a real gas the number is fixed. In helium II it depends on the temperature, in a way that can be calculated once the energy-momentum relation for quasi-particles is known. At absolute zero there are no quasi-particles, and their number increases as the temperature, and hence the energy of the fluid, is raised.

The relation between the energy and momentum of a quasi-particle [see top illustration on page 605] is of fundamental importance. Landau originally inferred it by fitting his theory to empirical data. Later Richard P. Feynman of the California Institute of Technology succeeded in calculating it approximately from a detailed analysis of the modes of motion.

Since the simple S-shaped curve contains the clue to superfluidity, it is worth a moment's consideration. Note the contrast between it and the energy-momentum curve for real particles, shown in gray. The latter curve starts horizontally, becoming steeper as it moves to the right. This simply reflects the fact that, as the velocity of a particle increases, its energy (which depends on the square of the velocity) goes up faster than its momentum (which depends on the first power of the velocity). The important point here is that the ratio of energy to momentum can have any value from zero upward. By contrast, the curve for quasi-particles rises sharply, at an angle to the horizontal. Hence the ratio of energy to momentum can never be less than some minimum value.

To see how the theory accounts for superfluidity, consider the simplest situation at absolute zero; then there are no quasi-particles, only pure background fluid. Imagine an object, say a golf ball, moving through the liquid. In an ordinary fluid the ball would quickly slow down, transferring its energy to the molecules. In the case of helium II the absorption of energy would mean the creation of a quasi-particle in the background fluid. If the golf ball is to slow

down, any energy it loses must be transferred to the quasi-particle. Now when the conservation of momentum is also taken into account, it is found that the ratio of the energy to the momentum acquired by the quasi-particle depends on the original speed of the ball. And if that speed is low enough, the ratio is less than the minimum possible value. Therefore a slow-moving ball cannot give up any energy to the fluid. By the same token, if the fluid is moving slowly, it cannot lose energy to the walls of the vessel. In short, it is superfluid.

Many other features of the strange behavior of helium II can be readily understood in terms of quasi-particles. Consider first the paradox of the two different viscosities. When helium II flows in a narrow channel, it is the background fluid (having the superfluid property of the liquid in its ground state) that gets through, while the quasi-particles do not. In the rotating cylinder viscometer, on the other hand, the gas of quasi-particles transmits a force from the outer to the inner cylinder even though the background fluid does not. The paradox is thus resolved because one is really dealing with two distinct entities: the background fluid, which is responsible for the flow through narrow openings; and the quasi-particles, which are responsible for the viscosity measured in the viscometer.

This idea of two interpenetrating fluids is strikingly supported by the following experiment. Helium II is placed in a container with an open bottom, which is then plugged with tightly packed emery powder [see top illustration on this page]. Background fluid can escape through the fine channels in the plug, but quasi-particles cannot. As the container empties, therefore, the concentration of quasi-particles in the remaining liquid is increased. But as we have said, the number of quasi-particles per unit volume is related to the temperature of the fluid. Hence the temperature of the liquid remaining in the container should also increase, and this is just what happens.

Since quasi-particles have momentum, they ought to produce observable mechanical effects. The Russian physicist Peter Kapitza demonstrated such an effect by heating helium II contained in an open bottle, which was in turn immersed in a bath of helium II. As the temperature rose, the number of quasi-particles in the bottle increased, thus "raising the pressure of the quasi-particle gas." As a result a jet of quasi-parti-

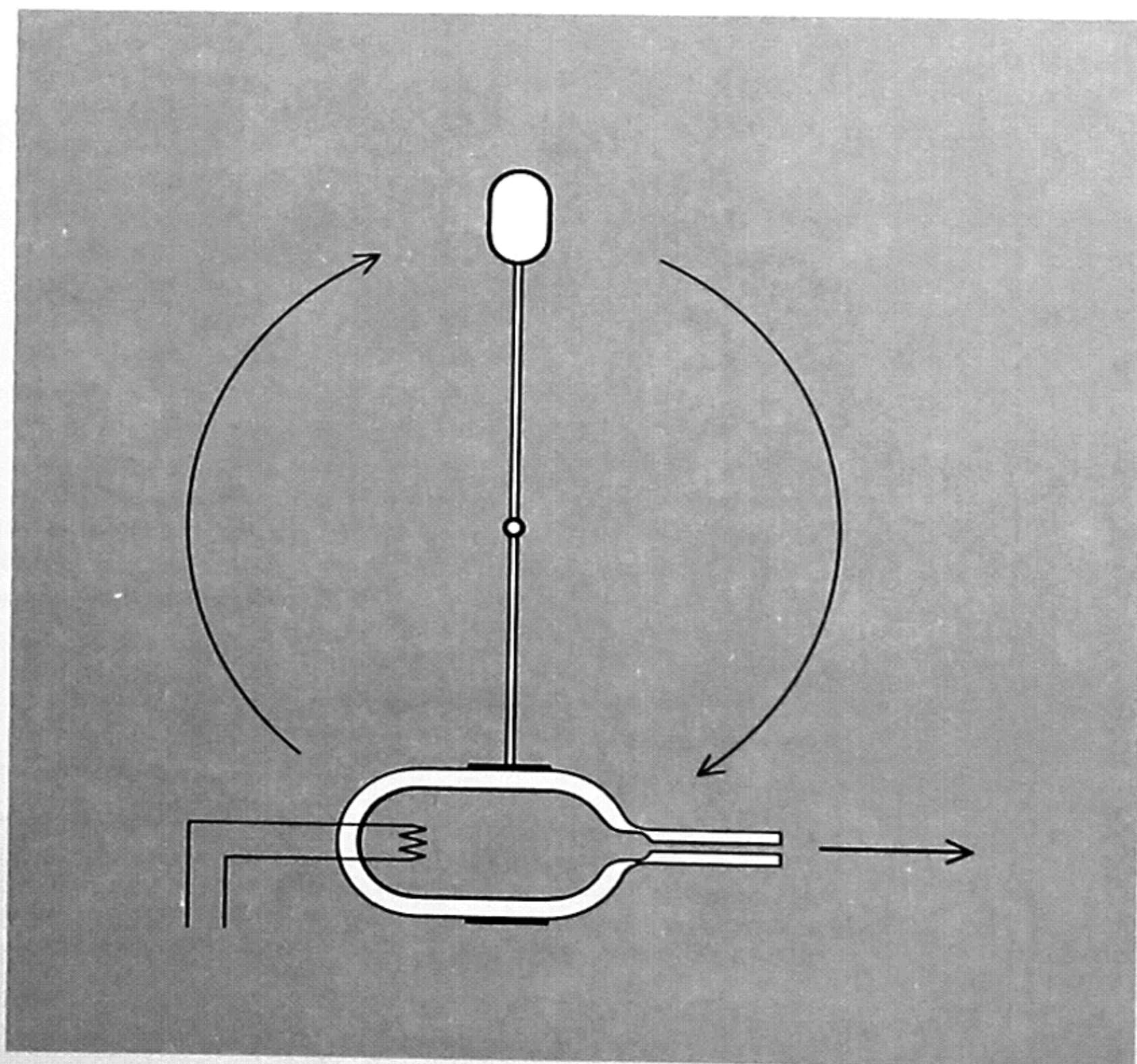
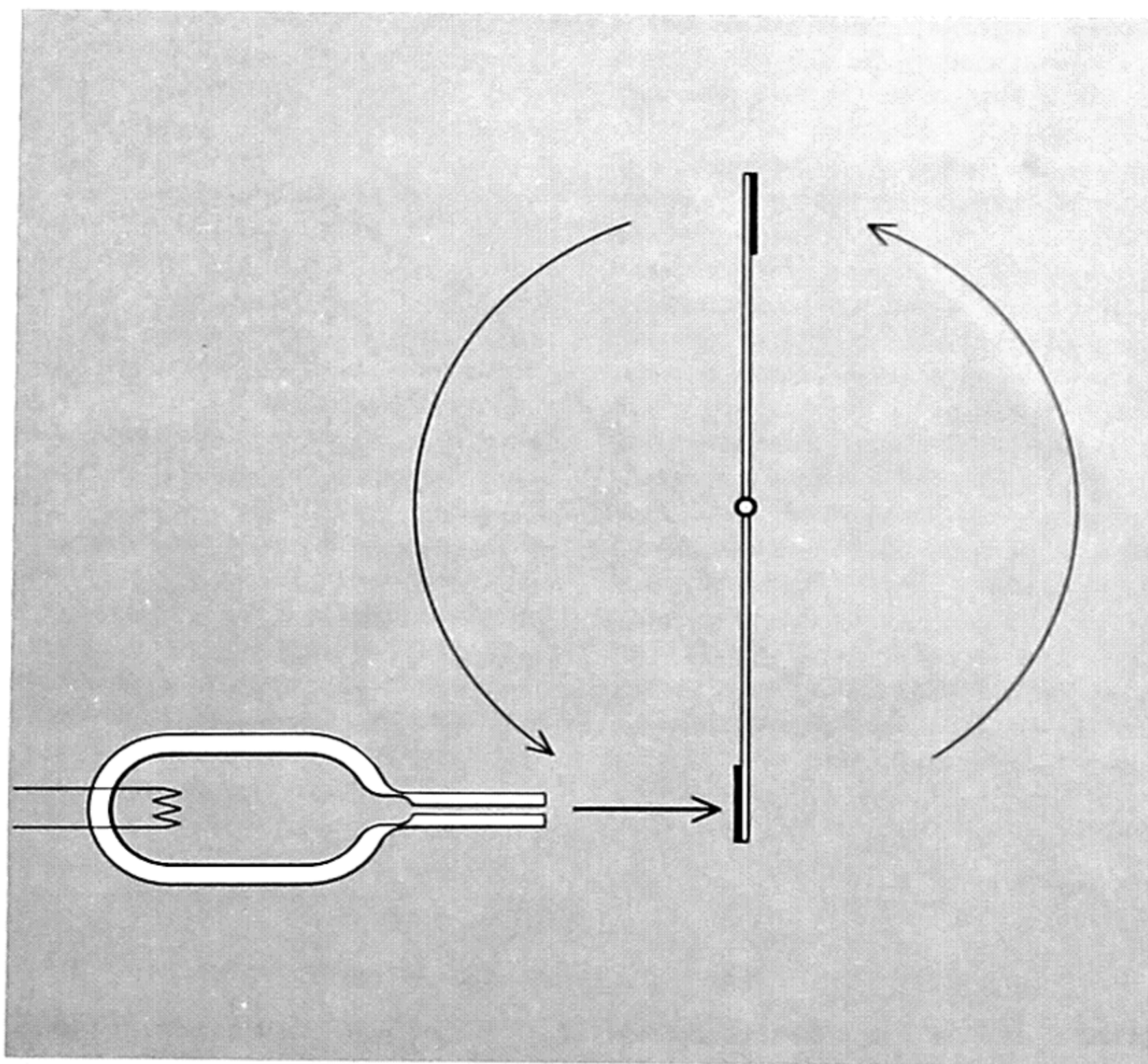
cles emerged from the bottle, deflecting a small vane placed in front of the mouth. When the bottle itself was suspended so that it was free to move, it recoiled when the heater was turned on.

A change in the setup of this experiment produces the fountain effect mentioned earlier. The bottle is partially immersed in the main helium II bath, and is open to it only through a very fine channel that passes background fluid but not quasi-particles. When the fluid in the bottle is heated, its level rises. As before, heating increases the pressure of the quasi-particle gas, but now it cannot escape. Instead it exerts a pressure on the background fluid which confines it and forces the level of liquid to rise in the bottle. In the process, background fluid passes from the main bath into the bottle.

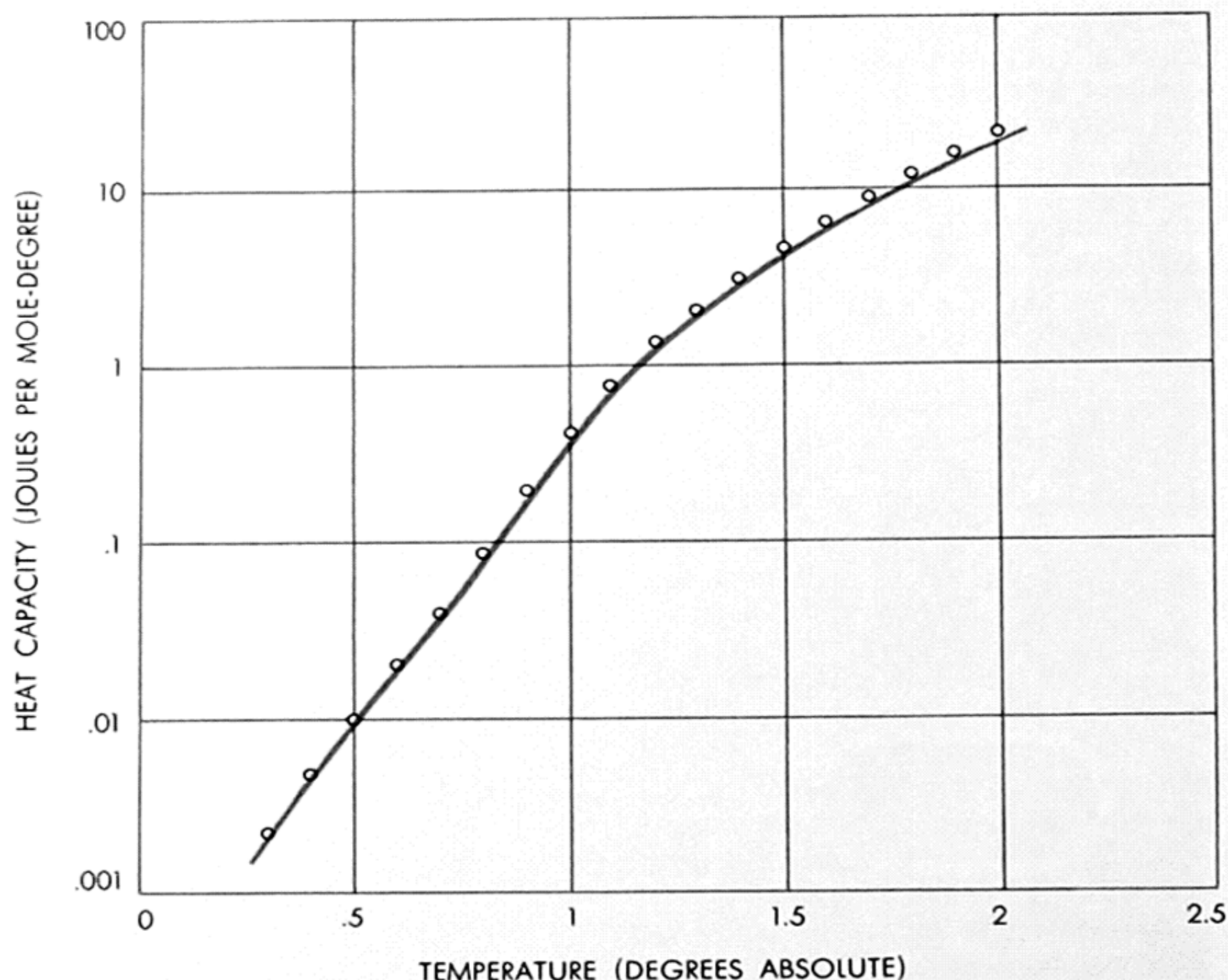
The phenomenon is completely analogous to osmotic pressure in ordinary liquids. If a bottle partly filled with sugar solution is connected to a bath of pure water through a membrane that is permeable to water but not to sugar molecules, water flows into the bottle, raising the level of the solution above that of the bath. In this experiment real particles (sugar molecules) take the place of quasi-particles, but the basic mechanism is the same.

Another demonstration of the relation between quasi-particles and temperature was devised by Kapitza's co-worker E. L. Andronikashvili. He suspended a stack of very closely spaced circular disks from a fiber, so that the stack could execute circular oscillations like those of the balance wheel of a watch. When the disks were immersed in helium II, the period (the time required for each oscillation) decreased rapidly as the temperature was reduced from the lambda point toward absolute zero. The period of course depends on the mass of the stack; the greater the mass, the longer the time. In helium II the background fluid passes readily between the disks, but the quasi-particles become clogged in the narrow spaces and are carried back and forth with the disks. Therefore they increase the effective mass of the stack and lengthen its period of oscillation. As the temperature is reduced, there are fewer quasi-particles and so the oscillation speeds up. This experiment can therefore provide a quantitative measurement of the mass that is contributed by quasi-particles at any temperature.

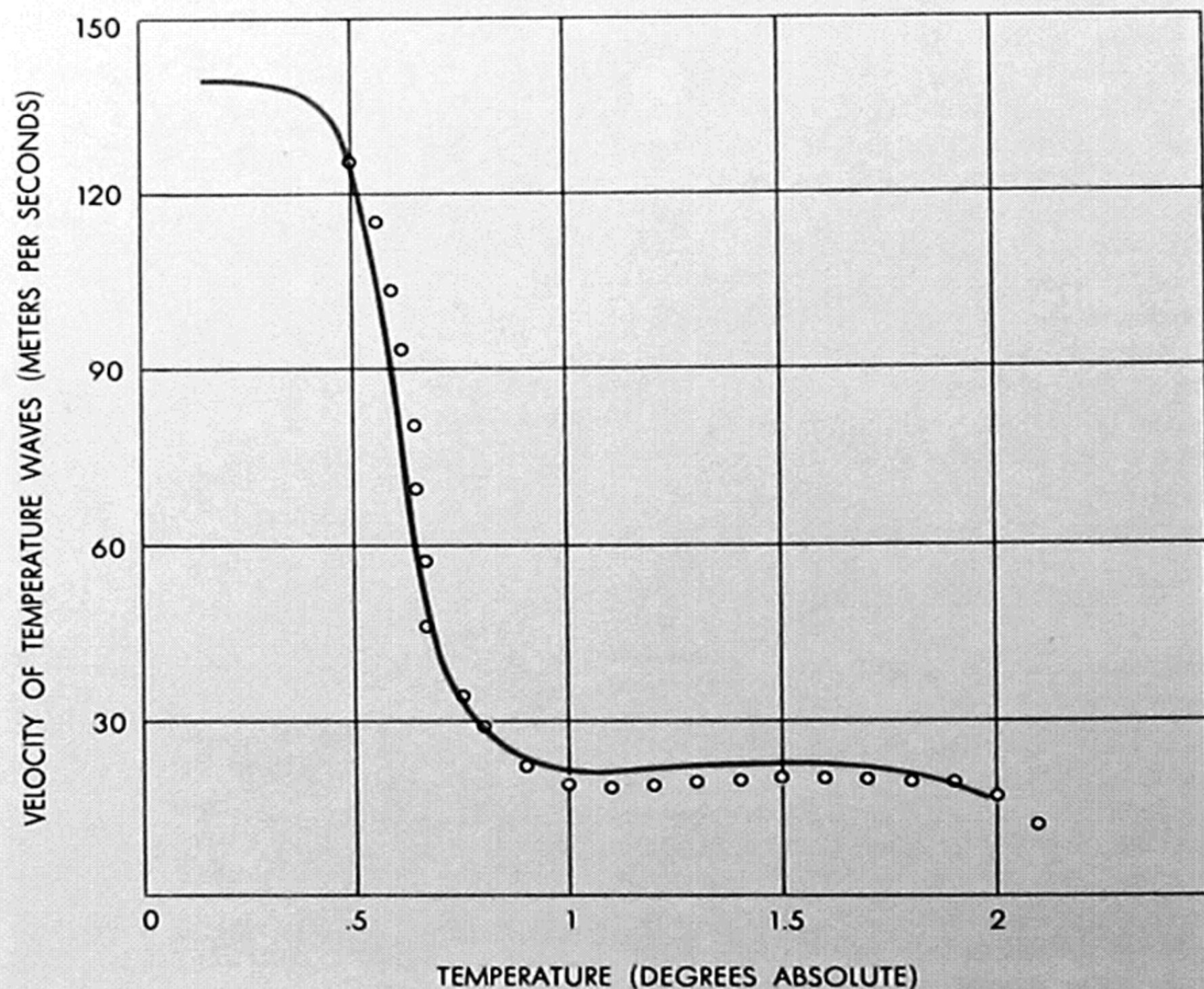
One of the most interesting features of the behavior of helium II was predicted by the quasi-particle theory before it was found in the laboratory. This



JET OF QUASI-PARTICLES emerges from a bottle of helium II when electric heater is turned on. Jet can be used to spin a small vane (*top*) or to make the bottle rotate (*bottom*).



HEAT CAPACITY of helium II (*colored curve*) was calculated from results of neutron-scattering experiments. Curve agrees with direct measurements of heat capacity (*circles*).



VELOCITY OF TEMPERATURE WAVES in helium II (*colored curve*) was also calculated from neutron-scattering experiments. Curve agrees with measured values (*circles*).

is the phenomenon of temperature waves, or, as it is sometimes called, second sound. An ordinary sound wave is set up in a fluid when a moving body such as the diaphragm of a loudspeaker produces a local compression of the material. This compression travels through the fluid at a well-defined speed, and can be detected at a distance from the source by its effect on a second diaphragm—for example, the eardrum. Sound travels through helium II just as through an ordinary fluid. When the liquid is compressed, the background fluid and the quasi-particle gas, both increase in density simultaneously. It is possible, however, to conceive of a local increase in density of the quasi-particle gas separately, the background fluid moving so as to keep the total density of the liquid constant. In that case the compression should travel through the quasi-particle gas with a definite speed, much as a sound wave moves through the air. But a local increase in the concentration of quasi-particles corresponds to a local increase in temperature of the liquid. Hence a local temperature increase in helium II should move through the liquid with all the properties of a wave motion. This wave motion has actually been set up by heating a sample of helium II in a localized region. The compression wave of quasi-particles can be detected some distance away with a thermometer.

All the foregoing experiments deal with macroscopic phenomena observed in helium II, and the theory of quasi-particles has certainly been very successful in accounting for the results. Very recently some experiments of a more microscopic kind have been carried out that subject the theory to a more detailed test. One was performed by L. Meyer and myself at the University of Chicago. We introduced a few electrically charged helium atoms (helium ions) into a sample of helium II and applied a voltage across the fluid. The ions move through the fluid under the influence of the electric force. We investigated the variation of the drift velocity with temperature, and found that the velocity increases very rapidly as the temperature is lowered below the lambda point. For example, at .6 degree absolute it is about 100,000 times greater than at the lambda point. Thus the superfluid character of helium II manifests itself also on an atomic scale. The ions are slowed by collisions with quasi-particles, and therefore speed up as the number of these quasi-particles decreases

with temperature. Note that collisions with individual helium atoms are of no significance to the motion of the ions. Indeed, in this experiment an ion can travel distances as great as 10,000 times the separation between atoms without being deflected by any of them. Again this is a typical quantum effect.

Finally let us turn to a very fundamental microscopic experiment, suggested by Feynman, in which a neutron beam is used to probe helium II. When neutrons are passed through any material, some of them are scattered in various directions by collisions with the atoms of the target. In the case of helium II at sufficiently low temperatures, neutrons of a suitable energy range are scattered by transferring some of their energy and momentum to the liquid as a whole. This means that a quasi-particle must be created having energy and momentum equal to that lost

by each neutron. The latter values can be determined from the known energy of the incoming neutrons and the measured energy of the deflected neutrons.

This experiment provides a direct measure of the energy and momentum of individual quasi-particles. The experimental curve obtained by John L. Yarnell and his colleagues at the Los Alamos Scientific Laboratory has exactly the shape previously discussed and constitutes a striking confirmation of Landau's theory. Furthermore, the curve provides a precise, quantitative measure of the energy-momentum relation for quasi-particles. It can therefore be used to compute many of the properties of helium II, such as heat capacity and the velocity of temperature waves, and to determine how these quantities vary with temperature. The calculated values agree very closely with direct measurements of these quantities.

Outside the realm of nuclear and high-energy physics, all systems, even biological organisms, can be said to be understood "in principle." That is, their atoms obey well-known laws of quantum mechanics and interact through well-known electromagnetic forces. Understanding "in principle," however, is meaningless when the systems are so complex that their behavior cannot be predicted from these laws and forces. Nor is the problem of complexity likely to be overcome simply by resorting to bigger and better electronic computers. More often the basic problem is to find a particular theoretical framework—a set of concepts well adapted to the system—which can facilitate and guide our thinking. Ideally these concepts should be derived from fundamental principles. In the case of superfluid liquid helium, the quasi-particle concept achieves precisely these aims.

The Author

F. REIF is associate professor of physics at the University of California. In 1939, at the age of 12, he left his native Austria, spent the first two years of World War II in France, and then came to the U. S. in 1941. Following undergraduate work at Columbia University he studied physics at Harvard University, where he acquired his Ph.D. in 1953. From 1953 until this September, when he joined the faculty at California, Reif was a joint member of the department of physics and the Institute for Metals at the University of Chicago. His interest in superfluidity is relatively recent; most of his research has been in solid-state physics and nuclear-magnetic resonance.

Bibliography

- APPLICATION OF QUANTUM MECHANICS TO LIQUID HELIUM. R. P. Feynman in *Progress in Low Temperature Physics*, Vol. 1, pages 17-53. Interscience Publishers, Inc., 1955.
- EXCITATIONS IN LIQUID HELIUM: NEUTRON SCATTERING MEASUREMENTS. J. L. Yarnell, G. P. Arnold, P. J. Bendt and E. C. Kerr in *The Physical Review*, Vol. 113, No. 6, pages 1,379-1,386; March 15, 1959.
- SCATTERING OF THERMAL ENERGY IONS IN SUPERFLUID LIQUID HE BY PHONONS AND He^3 ATOMS. Lothar Meyer and F. Reif in *Physical Review Letters*, Vol. 5, No. 1, pages 1-3; July 1, 1960.
- STUDY OF SUPERFLUIDITY IN LIQUID HE BY ION MOTION. F. Reif and L. Meyer in *The Physical Review*, Vol. 119, No. 4, pages 1, 164-1, 173; August 15, 1960.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

GRAVITY

by George Gamow

Albert Einstein showed that gravitation can be interpreted as a geometrical property of space-time. His further hope, of relating gravity and electromagnetism, is still unfulfilled.

In the days when civilized men believed that the world was flat they had no reason to think about gravity. There was "up" and "down." All material things tended naturally to move downward, or to fall, and no one thought to ask why. The notion of absolute up and down directions persisted into the Middle Ages, when it was still invoked to prove that the earth could not be round.

The first ray of light to pierce the mist of scholastic ideas about falling bodies issued from the work of Galileo Galilei. Since free fall was too fast to measure directly, Galileo decided to dilute the motion by studying bodies placed on an inclined plane. He argued—and at the time it was a novel argument—that since a ball resting on a horizontal surface does not move at all, and since a ball falling parallel to a vertical surface moves as fast as it would if the surface were not there, a ball on an inclined surface should roll with an intermediate speed depending on the angle of inclination. Letting balls roll down planes tilted at various angles, he observed their rates of travel and the distances covered in different time intervals, which he measured with a water clock. The experiments showed that at any angle the speed increases in direct proportion to time (counted from the moment of release) and that the distance covered increases in proportion to the square of the time. Galileo also observed that a massive iron ball and a much lighter wooden ball roll down side by side if released simultaneously from the same height on the same inclined plane.

As another way to dilute free fall he employed simple pendulums—weights suspended by thin strings. Here the steepness of the arc along which the weight travels is adjusted by changing the length of the string. Pendulums of the same length proved to have the same

period of oscillation even when the weight was varied, a result in agreement with the outcome of the inclined-plane experiments. From all these observations Galileo was led to infer that in free fall all material bodies, light or heavy, also move in exactly the same way. This idea directly contradicted the opinion of the then prevailing Aristotelian school of philosophy, which held that heavier bodies fall faster than light ones. According to the celebrated legend, which may or may not be true, Galileo climbed the leaning tower of Pisa and dropped a light and a heavy ball, which hit the ground simultaneously, to the consternation of contemporary philosophers.

Newton's Law of Gravity

These studies laid the foundation for the science of mechanics. The main structure was erected by Isaac Newton, who was born the year Galileo died. With his laws of motion Newton introduced the notions of force and of inertial mass. When a force is applied to material bodies, it changes their speed or direction of motion or both. Their inertial mass opposes these changes. Newton stated that the rate of change of velocity (acceleration) of an object is directly proportional to the force acting on it and inversely proportional to its mass. Doubling the force doubles the acceleration; doubling the mass cuts the acceleration in half; if both force and mass are doubled, the acceleration is unchanged.

In the light of this law Galileo's conclusion about free-falling bodies implies a fact that is usually taken for granted, but which is actually very curious; namely, the weight of a body (that is, the gravitational pull of the earth upon it) is strictly proportional to its inertial mass. Otherwise an iron and a wooden ball of the same size would not fall at

the same rate. If the two objects have the same acceleration when they are dropped, the inertial mass opposing a change of motion in the iron ball must be greater than that in the wooden ball in exactly the same proportion that the downward force on the iron ball is greater. This proportionality is far from trivial; in fact, it holds true only for gravity and not for other familiar forces such as those of electricity and magnetism. Thus while an electron and a proton would fall with equal acceleration in a gravitational field, when these particles are placed in an electric field the electron is accelerated 1,836 times faster.

From his analysis of balls (or apples) that fall toward the earth Newton went on to consider gravitation in wider terms. His line of thought is demonstrated by a very interesting discussion in his *Principia*. Suppose, he said, we shoot a bullet horizontally from the top of a mountain so high that it rises above the atmosphere [see illustration on page 612]. The bullet will follow a curved trajectory and hit the earth's surface some distance away from the base of the mountain. The greater the muzzle velocity, the farther away from the mountain the bullet will land. At a sufficiently high initial velocity the bullet will come to earth at a point directly opposite the mountain; at still higher velocity it will never hit the ground but will continue to revolve around the earth like a little moon. If, Newton argued, it is possible in this way to make an artificial satellite, why not assume that the motion of the natural moon is also a free fall? And if the moon revolves around the earth because of the earth's gravitational attraction, is it not logical to assume that the earth itself is held in orbit around the sun by the force of the sun's gravity? Then is this not also true for all the other

planets and their satellites? So originated the profoundly important idea of universal gravitation, which states that all material bodies in the universe attract one another with forces determined by their masses and mutual distances.

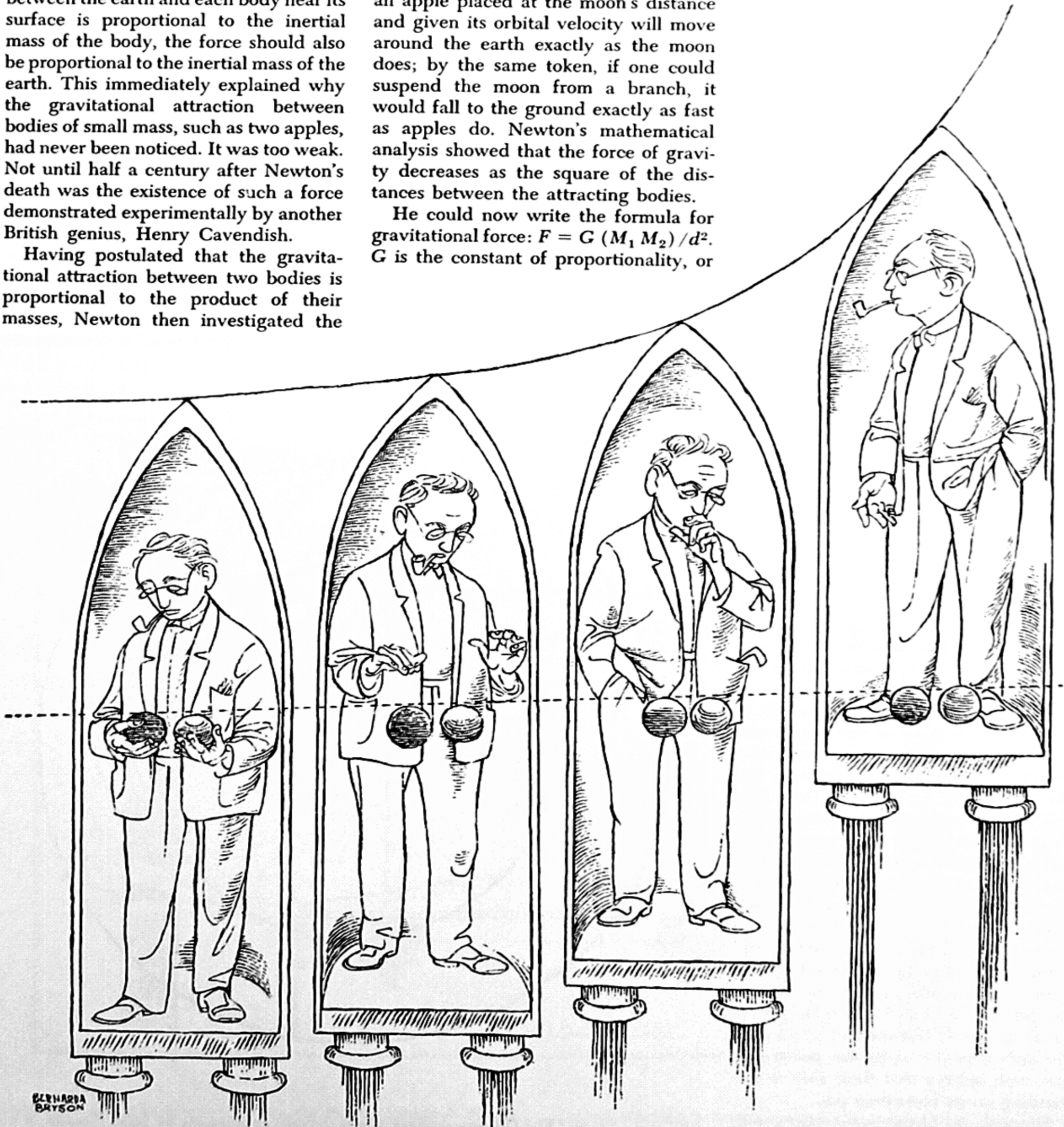
To establish the exact relation of force to mass and distance, Newton began by assuming that, since the force between the earth and each body near its surface is proportional to the inertial mass of the body, the force should also be proportional to the inertial mass of the earth. This immediately explained why the gravitational attraction between bodies of small mass, such as two apples, had never been noticed. It was too weak. Not until half a century after Newton's death was the existence of such a force demonstrated experimentally by another British genius, Henry Cavendish.

Having postulated that the gravitational attraction between two bodies is proportional to the product of their masses, Newton then investigated the

dependence on distance. He compared the force necessary to hold the moon in its orbit at the distance of 60 earth radii with the force on an apple at the distance of only one radius from the center of the earth. It is important to realize here that the great difference in mass between the two bodies does not affect the validity of the comparison. As a matter of fact, an apple placed at the moon's distance and given its orbital velocity will move around the earth exactly as the moon does; by the same token, if one could suspend the moon from a branch, it would fall to the ground exactly as fast as apples do. Newton's mathematical analysis showed that the force of gravity decreases as the square of the distances between the attracting bodies.

He could now write the formula for gravitational force: $F = G (M_1 M_2) / d^2$. G is the constant of proportionality, or

the gravitational constant. It is a very small number; if the masses are measured in grams and the distance in centimeters, G is approximately .000000066. This means that a pair of one-gram weights separated by one centimeter attract each other with a force a little



PRINCIPLE OF EQUIVALENCE enunciated by Einstein states that accelerated motion produces effects indistinguishable from those of a gravitational field. If an observer in a uniformly accelerating spaceship simultaneously releases two balls of different

weight, he will see them fall toward the floor at the same rate. An outside observer would say that the balls continue to move upward (*broken line*) with the speed of the ship at the moment of release, while the floor, moving up at an accelerating rate, overtakes them.

more than six hundred-millionths of a dyne, or about six hundred-billionths of the weight of a gram.

Combining the law of gravitation with his laws of motion, Newton was able to derive mathematically the rules governing planetary motion that had been discovered by Johannes Kepler. In the memorable era that followed, Newton and his successors explained the motions of celestial bodies down to the most minute details. But the nature of gravitational interaction, and in particular the reason for the mysterious proportionality between gravitational mass and inertial mass, remained completely hidden for more than 200 years.

Einstein's Law of Gravity

Then, in 1914, Albert Einstein lifted the veil. The ideas he put forward grew out of his formulation of the special theory of relativity a decade earlier. That theory is based on the postulate that no observations made inside an enclosed chamber can answer the question of whether the chamber is at rest or moving along a straight line at constant speed. Thus a person in the situation of the author as he writes these lines—in an inside cabin of the S.S. *Queen Elizabeth* sailing on a smooth sea—can perform no experiment, mechanical, optical or any other kind, that will tell him whether the ship is really moving or still in port. But let a storm come up and the situation changes painfully; the deviation from uniform motion is all too apparent.

In order to deal with the problem of nonuniform motion Einstein imagined a laboratory in a spaceship located far from any large gravitating masses. If the vehicle is at rest, or in uniform motion with respect to distant stars, the observers inside, and all their instruments that are not secured to the walls, will float freely. There will be no up and no down. As soon as the rocket motors are started and the ship accelerates, however, instruments and people will be pressed to the wall opposite the direction of motion. This wall will become the floor, the opposite wall will become the ceiling and the people will be able to stand up and move about much as they do on the ground. In fact, if the acceleration is equal to the acceleration of gravity on the surface of the earth, the passengers may well believe that their ship is still standing on its launching pad.

Suppose one of the passengers simultaneously releases two spheres, one of iron and one of wood, which he has been holding next to each other in his hands.

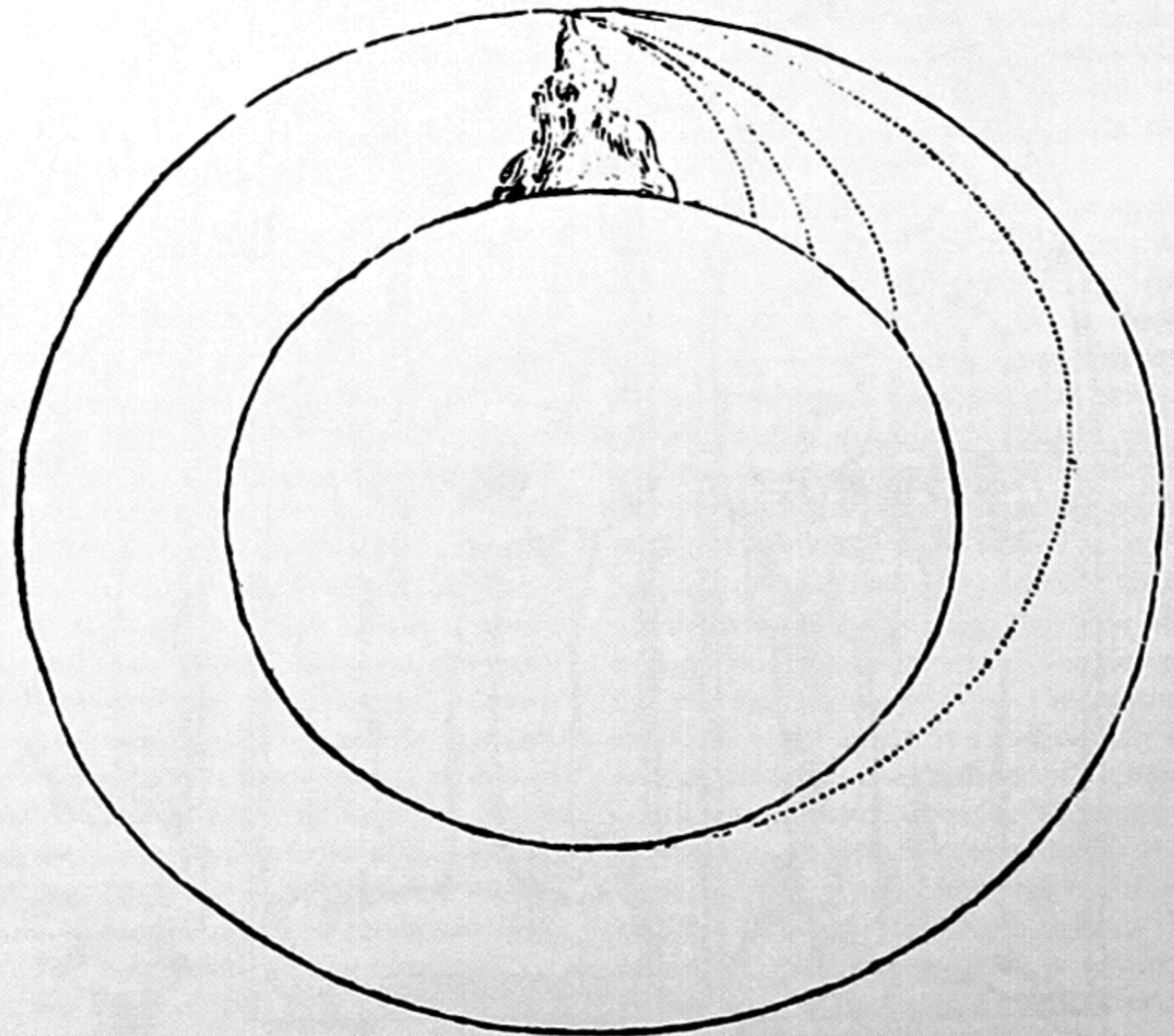
What "actually" happens can be described as follows: While the spheres were held they were undergoing accelerated motion, along with the observer and the whole ship. When they are released, they are no longer driven by the rocket engines. Now they will move side by side, each with a velocity equal to that of the spaceship at the moment of release. The ship itself, however, will continuously gain speed and the "floor" of the ship will quickly overtake the two spheres and hit them simultaneously.

To the observer inside the ship the experiment will look different. He will see the balls drop and hit the "floor" at the same time. Recalling Galileo's demonstration from the leaning tower of Pisa, he will be persuaded that an ordinary gravitational field exists in his space laboratory.

Both descriptions of the observed event are correct; the equivalence of the two points of view is the foundation of Einstein's relativistic theory of gravity. This so-called principle of equivalence

between observations carried out in an accelerated chamber and in a "real" gravitational field would be trivial, however, if it applied only to mechanical phenomena. Einstein's deep insight was that the principle is quite general and holds also for optical and other electromagnetic phenomena.

Imagine a beam of light propagating across the space laboratory in a "horizontal" direction. Its path can be traced by means of a series of vertical fluorescent glass plates spaced at equal distances [see illustration on page 614]. Again what actually happens is that the beam travels in a straight line at constant speed, while the glass plates move across its path at an ever increasing speed. The beam takes the same time to travel from each plate to the next, but the plates move farther during each successive interval. Hence the pattern of fluorescent spots shows the floor approaching the light beam at an increasing rate. If the observer inside the chamber draws a line through the spots, it will look to him like



ARTIFICIAL SATELLITE was envisaged as a thought experiment by Isaac Newton in his *Principia*, from which this diagram is reproduced. Bullet fired horizontally from a mountaintop falls farther from base as its muzzle velocity is increased. At sufficiently high speed it circles the earth, suggesting that the moon is also falling in the earth's gravitational field.

a parabola bending toward the floor. Since he considers acceleration phenomena as being caused by gravity, he will say that a light ray is bent when propagating through a gravitational field.

Thus, concluded Einstein, if the principle of equivalence holds in all of physics, light rays from distant stars that pass close to the sun on their way to the earth should bend toward the sun. This prediction was brilliantly confirmed in 1919 by a party of British astronomers observing a total solar eclipse in Africa. With the obscuring sunlight extinguished by the moon, stars near the edge of the solar disk were seen to be displaced about 1.75 seconds of arc away from the sun.

Relativistic Merry-Go-Round

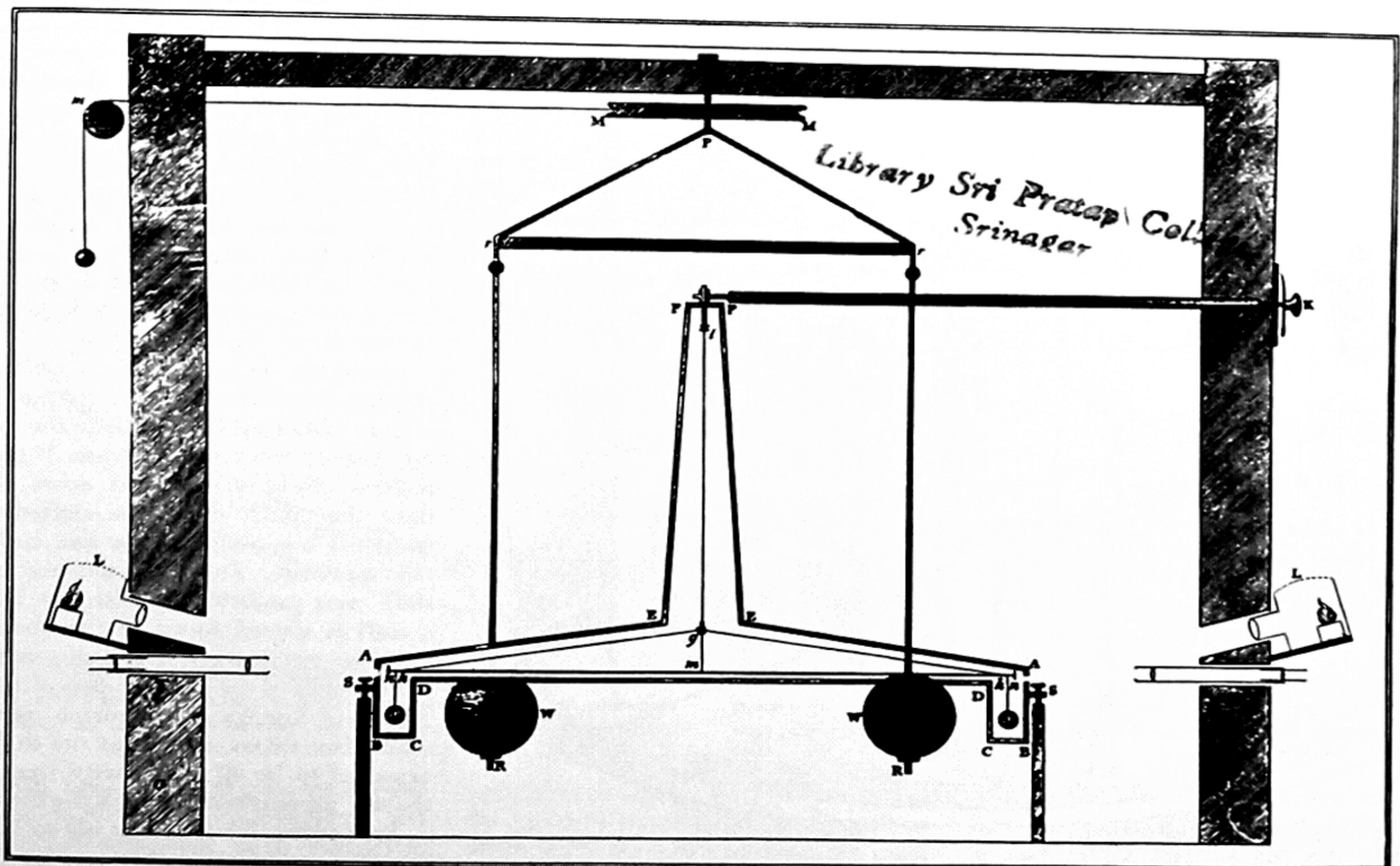
Let us next consider another type of accelerated motion—uniform rotation. (A body moving at constant speed on a circular path is accelerated because of its continuous change of direction.) Im-

agine a merry-go-round with a curtain around it so that people inside cannot tell by looking at the surroundings that it is rotating. If the merry-go-round is turning, the observers will be aware of centrifugal force, which pushes them out toward the rim. A ball placed on the platform will roll away from the center. The centrifugal force acting on any object on the platform will be proportional to the inertial mass of the object, so that here again the effect of accelerated motion can be considered as equivalent to that of a gravitational field. It is a peculiar field, to be sure; it is quite different from the field on the surface of the earth or of any other spherical body. The force is directed away from the center of the system, not toward it; and instead of decreasing as the square of the distance from the center, it increases proportionately to that distance. Moreover, the field has cylindrical symmetry around a central axis rather than spherical symmetry around a central point. Nevertheless, the equivalence principle

holds, and the field can be interpreted as being caused by gravitating mass distributed at large distances all around the symmetry axis.

How will light propagate through this field? Suppose a light source that sends out rays in all directions is located at a point, *A*, on the periphery of the rotating disk, and is observed at a second point, *B*, also on the periphery. According to the basic law of optics, light always propagates along the shortest path. But what is the shortest path between *A* and *B*? To measure the length of various lines connecting the points *A* and *B* the observer uses the old-fashioned but always safe method of counting the number of yardsticks that can be placed end to end along the line [see illustration on page 615].

As we watch the experiment from outside, we recall the special theory of relativity, which tells us that moving yardsticks shrink in the direction of their motion. Therefore we see that if the observer measures along the "true" straight



TORSION BALANCE was used by the British physicist Henry Cavendish to measure gravitational attraction between small masses. In this diagram, reproduced from his paper in the *Philosophical Transactions of the Royal Society*, two-inch lead spheres

(*x*) are attached to beam suspended by torsion wire (*lg*). Twelve-inch spheres (*W*) are placed so as to twist the beam first in one direction and then in the other. Turning is observed through system of lenses at each side. Total rotation is measure of attractive force.

line from A to B, his sticks will contract and he will need more of them to measure that line than if the platform were not moving. Now an interesting point arises. The closer a yardstick is to the center of the merry-go-round, the less its linear velocity and therefore the smaller

its contraction. By bending the line of yardsticks toward the center the observer decreases the number he needs to go from A to B. Although the "actual" distance is somewhat longer, the increase is more than compensated for by the smaller shrinkage of each yardstick. A

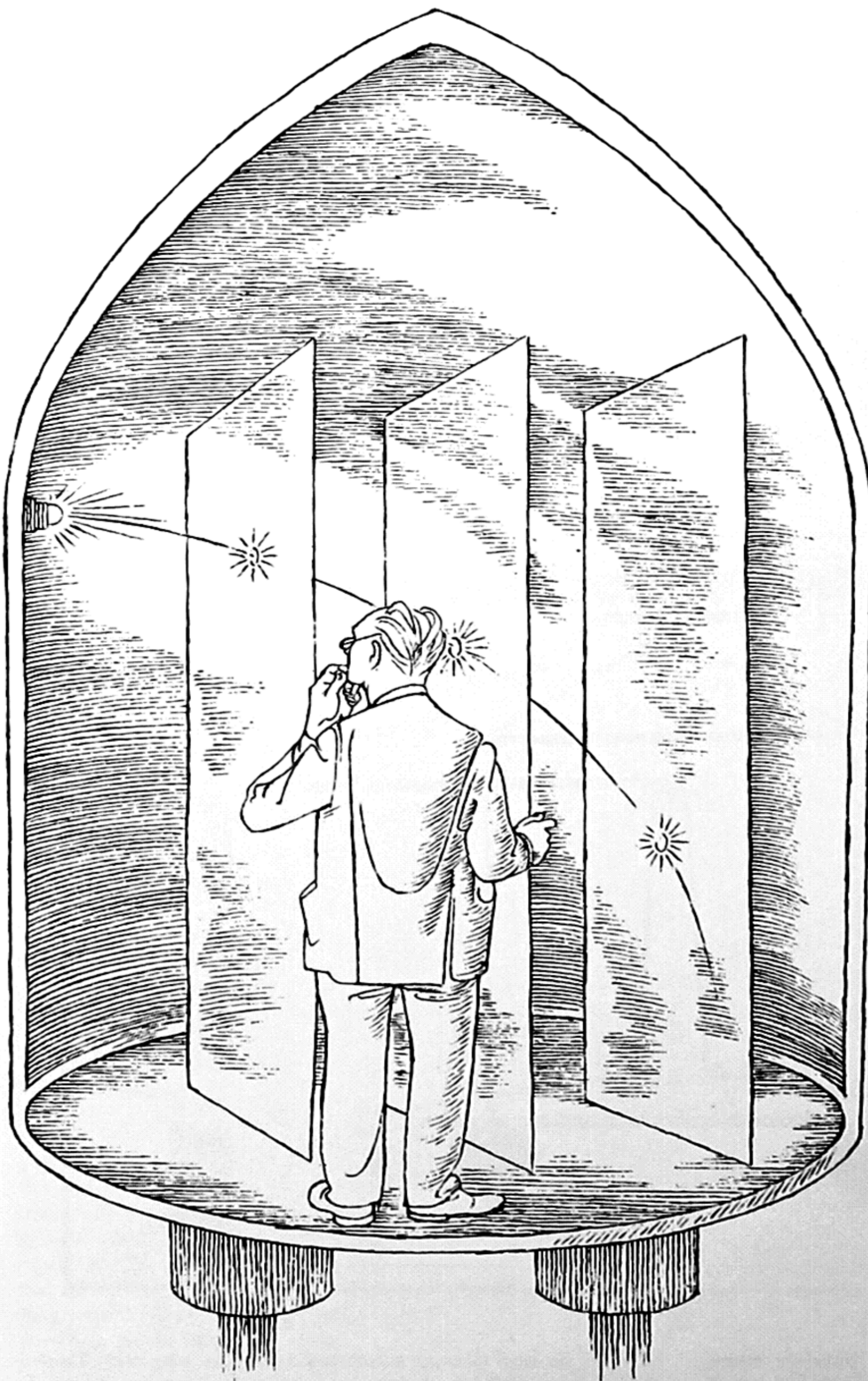
light ray following this shortest path, heading inward at the start of its journey and then bending outward, can be considered to be deflected by the apparent gravitational field, which is directed radially outward.

Before leaving the merry-go-round let us consider one more experiment. A pair of identical clocks are placed on the platform, one near the center and the other at the edge. As in the case of the yardsticks, the outer clock is moving faster than the inner one, and again special relativity predicts a difference in their behavior. In addition to causing yardsticks to contract, motion makes clocks run slow. Therefore the outer clock will lose time with respect to the inner one. Now the observer who interprets the acceleration effects in terms of a gravitational field will say that the clock placed in the higher gravitational potential (that is, in the direction in which gravitational force acts) runs slower.

Although we cannot go into details here, Einstein's argument shows that the same effect is expected in a normal gravitational field such as that on the earth. Here the field is directed downward, so that a clock at sea level runs slower than one on top of a mountain. The slowing down applies equally to all other physical, chemical and biological phenomena, and a typist working on the first floor of the Empire State Building will age slower than her twin sister working on the top floor. Stronger fields produce greater retardation. A clock on the surface of the sun would run .0001 per cent slower than a terrestrial clock.

Obviously we cannot put a clock on the sun, but we can watch the rate of atomic vibrations that produce the various lines in the solar spectrum. If these natural clocks are slowed down, the light they emit should be shifted toward the low-frequency, or red, end of the spectrum. This "gravitational red shift" was predicted by Einstein. Such a shift is indeed found in the lines of the solar spectrum, but it is so small as to be almost at the limit of observational precision. Spectra of the much denser white-dwarf stars, where the red shift is expected to be 40 times larger than on the sun, agree quite well with the theory.

Astronomical evidence is not so satisfying as experiments that can be performed in a terrestrial laboratory. Until a couple of years ago, however, there seemed to be no hope of measuring the minute difference predicted between clocks at different heights in the earth's gravitational field. Then R. L. Mössbauer, working at the University of



CURVATURE OF LIGHT is detected by observer in accelerating rocket. To an observer outside, the light beam travels along a straight, horizontal path and crosses each successive plate of glass at a point nearer the floor because of the upward acceleration of the plates.



CENTRIFUGAL FIELD OF FORCE, such as the one on a rotating merry-go-round, can also be interpreted in terms of gravitating mass. As explained in the text, an observer on the platform

would find that a curved line is the shortest distance between points on the periphery. Since light rays travel along the shortest path, or geodesic, they are expected to curve in this type of field.

Munich, found a way to produce nuclear gamma rays of very pure frequency and to measure extremely small changes in their frequency [see "The Mössbauer Effect," by Sergio De Benedetti; *SCIENTIFIC AMERICAN* Offprint 271]. Seizing on the new opportunity, several workers proceeded to show that two nuclear "clocks" separated by only a few tens of feet in the earth's field run at measurably different rates, and the difference is ex-

actly that predicted by Einstein, within the limits of experimental error. Still another verification, if any more are needed, will almost certainly be obtained when an atomic clock in an artificial satellite is compared with one on the ground.

So we see that in a gravitational field clocks run slow, light rays bend in the direction of the field and a straight line is not the shortest distance between two

points. Yet how can one define "straight line" other than as the path of light in a vacuum, or the shortest distance between two points? Einstein's idea was to retain this definition. Instead of saying that light rays and shortest distances are curved, he suggested that space itself (more accurately space-time) is curved. It is difficult to conceive of a curved three-dimensional space, let alone a curved four-dimensional space-time, but

some idea of what it means can be gained from an analogy with two-dimensional surfaces. The Euclidean geometry we all learned at school pertains to figures that can be drawn on a plane. If geometrical figures are drawn on curved surfaces, for example a sphere or a surface shaped like a saddle [see illustration on this page], many of the Euclidean theorems do not hold.

In particular, the sum of the angles of a plane triangle is equal to 180 degrees. In a spherical triangle the sum of the angles is greater than 180 degrees, and in a triangle drawn on a saddle surface it is less. True, the lines forming triangles on spherical and saddle surfaces are not straight from the three-dimensional point of view, but they are the "straightest" (i.e., shortest) lines between the points if one is confined to the surface in question. Mathematicians call such lines geodesic lines, or simply geodesics.

In three-dimensional space a geodesic line is by definition the path along which a light ray would propagate. Consider a triangle formed by three such geodesics. If the sum of the angles is equal to 180 degrees, the space is said to be flat. If the sum is more than 180 degrees, we say that the space is spherelike, or positively curved; if it is less than 180 degrees, we say that it is saddle-like, or negatively curved. Because of the bending of light toward the sun, astronomers located on earth, Mars and Venus would

measure more than 180 degrees in the angles of the triangle formed by light rays traveling between the planets [see illustration on page 617]. Hence we can say that the space around the sun is positively curved. On the other hand, in the merry-go-round type of gravitational field, the sum of angles of a triangle is less than 180 degrees, and this space is curved in the negative sense.

The foregoing arguments represent the foundation of Einstein's theory of gravity. In the Newtonian view the sun produces in the space around it a field of force that makes the planets move along curved trajectories instead of straight lines. In Einstein's picture space itself becomes curved and the planets move along the straightest (geodesic) lines in that curved space. Here we are speaking of geodesics in the four-dimensional space-time continuum [see illustration on page 618]. It would, of course, be wrong to say that the orbits themselves are geodesic lines in three-dimensional space.

Einstein's interpretation of gravity as the curvature of space-time does not lead to exactly the same results as those of the classical Newtonian theory. We have already mentioned the bending of light. The relativistic theory also gives slightly different answers for the motions of material bodies. For example, it explained the difference between the calculated and observed rates of precession of the major axis of Mercury's orbit,

which represented a long-standing mystery of classical celestial mechanics.

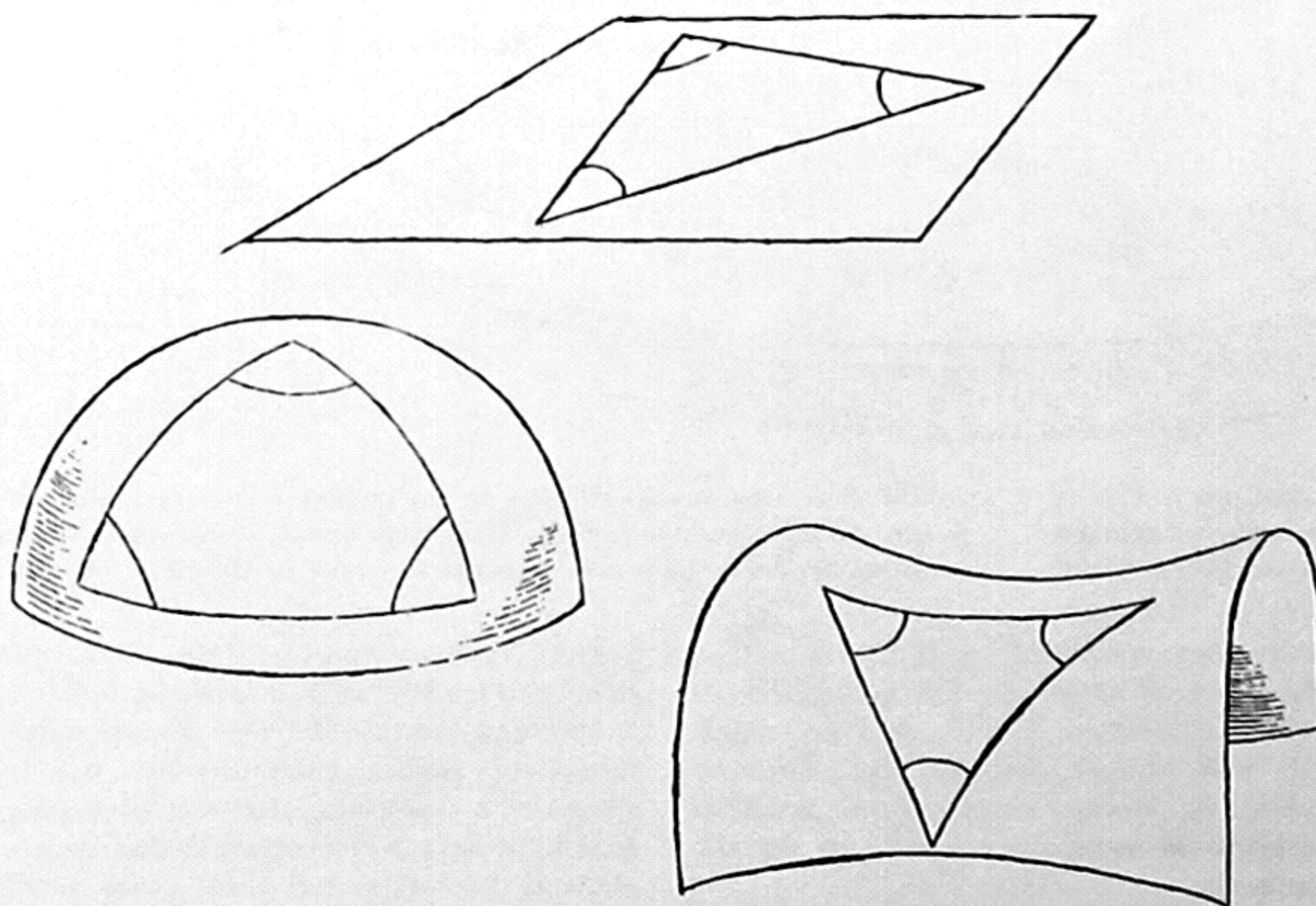
Gravity Waves

Newton's law of gravitational interaction between masses is quite similar to the law of electrostatic interaction between charges, and Einstein's theory of the gravitational field has many common elements with James Clerk Maxwell's theory of the electromagnetic field. So it is natural to expect that an oscillating mass should give rise to gravitational waves just as an oscillating electric charge produces electromagnetic waves. In a famous article published in 1918 Einstein indeed obtained solutions of his basic equation of general relativity that represent such gravitational disturbances propagating through space with the velocity of light. If they exist, gravitational waves must carry energy; but their intensity, or the amount of energy they transport, is extremely small. For example, the earth, in its orbital motion around the sun, should emit about .001 watt, which would result in its falling a millionth of a centimeter toward the sun in a billion years!

No one has yet thought of a way to detect waves so weak. In fact, some theorists, among them Sir Arthur Eddington, have suggested that gravitational waves do not represent any physical reality at all but are simply a mathematical fiction that can be eliminated from the equation by a suitable choice of space-time co-ordinates. More thorough analysis indicates, however, that this is not the case and that gravitational waves, weak though they may be, are real.

Are gravitational waves divided into discrete energy packets, or quanta, as electromagnetic waves are? This question, which is as old as the quantum theory, was finally answered two years ago by the British physicist P. A. M. Dirac. He succeeded in quantizing the gravitational-field equation and showed that the energy of gravity quanta, or "gravitons," is equal to Planck's constant, h , times their frequency—the same expression that gives the energy of light quanta or photons. The spin of the graviton, however, is twice the spin of the photon.

Because of their weakness gravitational waves are of no importance in celestial mechanics. But might not gravitons play some role in the physics of elementary particles? These ultimate bits of matter interact in a variety of ways, by means of the emission or absorption of appropriate "field quanta." Thus electromag-



SPACE CURVATURE is illustrated in two dimensions. In flat space (top) the angles of a triangle total 180 degrees; in spherelike or positively curved space (middle), more than 180 degrees; in saddle-like or negatively curved space (bottom), less than 180 degrees.

netic interactions (for example the attraction of oppositely charged bodies) involve the emission or absorption of photons; presumably gravitational interactions are similarly related to gravitons. In the past few years it has become clear that the interactions of matter fall into distinct classes: (1) strong interactions, which include electromagnetic forces; (2) weak interactions such as the "beta decay" of a radioactive nucleus, in which an electron and a neutrino are emitted; (3) gravitational interactions, which are vastly weaker than the ones called "weak."

The strength of an interaction is related to the rate, or probability, of the emission or absorption of its quantum. For example, a nucleus takes about 10^{-12} second (a millionth of a billionth of a second) to emit a photon. In comparison the beta decay of a neutron takes 12 minutes—about 10^{14} times longer. It can be calculated that the time necessary for the emission of a graviton by a nucleus is 10^{60} seconds, or 10^{53} years! This is slower than the weak interaction by a factor of 10^{58} .

Now, neutrinos are themselves particles with an extremely low probability of absorption, that is, interaction, with other types of matter [see "The Neutrino," by Philip Morrison; SCIENTIFIC AMERICAN Offprint 230]. They have no charge and no mass. As long ago as 1933 Niels Bohr inquired: "What is the difference between [neutrinos] and the quanta of gravitational waves?" In the so-called weak interactions neutrinos are emitted together with other particles. What about processes involving only neutrinos—say, the emission of a neutrino-antineutrino pair by an excited nucleus? No one has detected such events, but they may occur, perhaps on the same time scale as the gravitational interaction. A pair of neutrinos would furnish a spin of two, the value calculated for the graviton by Dirac. All this is, of course, the sheerest speculation, but a connection between neutrinos and gravity is an exciting theoretical possibility.

Gravity and Electromagnetism

In the laboratory diary of Michael Faraday appears the following entry in 1849: "Gravity. Surely this force must be capable of an experimental relation to electricity, magnetism and other forces, so as to build it up with them in reciprocal action and equivalent effect. Consider for a moment how to set about touching this matter by facts and trial." The numerous experiments he undertook to discover such a relation were

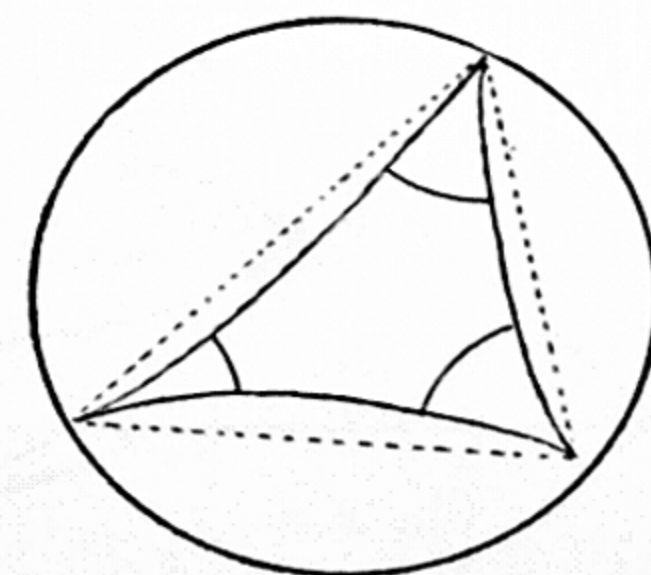
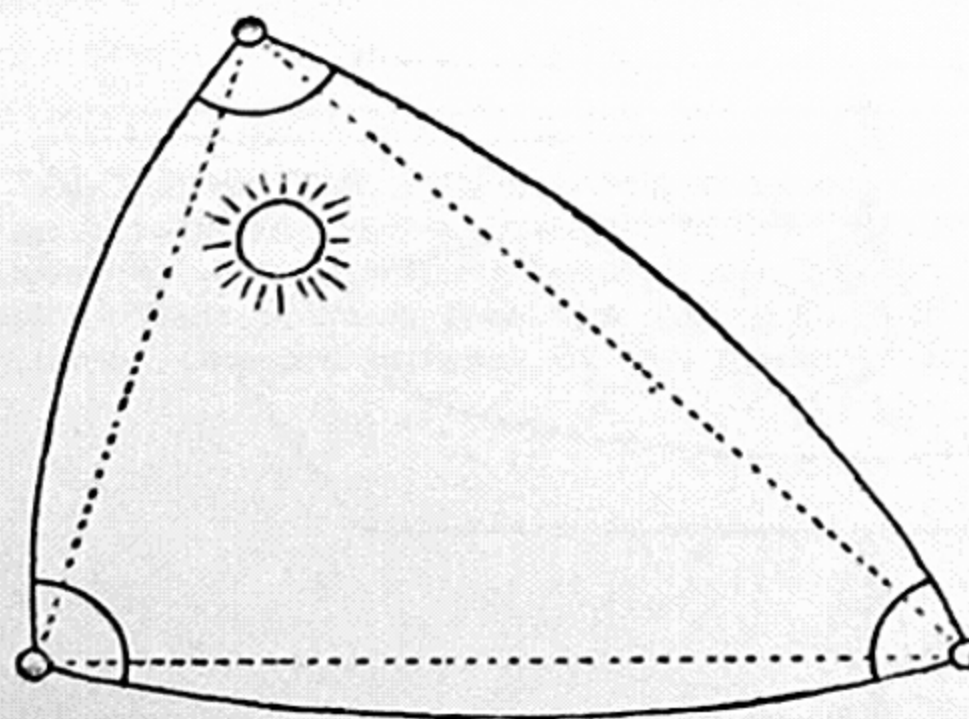
fruitless, and he concludes that part of his diary with the words: "Here end my trials for the present. The results are negative. They do not shake my strong feeling of the existence of a relation between gravity and electricity, though they give no proof that such a relation exists." Subsequent experimental efforts have not been any more successful.

A theoretical attack aimed at bringing the electromagnetic field into line with the gravitational field was undertaken by Einstein. Having reduced gravity to the geometrical properties of a space-time continuum, he became convinced that the electromagnetic field must also have some purely geometrical interpretation. However, the "unified field" theory, which grew out of this conviction, had hard going, and Einstein died without producing anything so simple, elegant and convincing as his earlier work. Today fewer and fewer physicists are working at unified-field theory; most are persuaded that the effort to geometrize the electromagnetic field is futile. It seems, at least to the author, that the true relation between gravitational and electromagnetic forces is to be found only through an understanding of the nature of elementary particles—an understanding of why there exist particles with just certain inertial masses and not others—and of the relation between the masses and the electric and magnetic properties of the particles.

As a sample of one of the basic questions in this field, consider again the relative strength of gravitational and electromagnetic interactions. Instead of comparing the times required for emission of quanta, let us compare the actual strength of the electrostatic and gravitational forces between a pair of middleweight particles, say pi mesons. Computation shows that the ratio of electrostatic to gravitational force equals the

square of the charge on an electron divided by the square of the mass of the particles times the gravitational constant: $e^2 / M^2 G$. For two pi mesons the value is 10^{40} . Any theory that claims to describe the relation between electromagnetism and gravity must explain this ratio. It should be pointed out that the ratio is a pure number, one that remains unchanged no matter what system of units is used for measuring the various physical quantities. Such dimensionless constants, which can be derived in a purely mathematical way, often turn up in theoretical formulas, but they are usually small numbers such as 2π , $5/3$ and the like.

How can one derive mathematically a constant as large as 10^{40} ? Some 20 years ago Dirac made an interesting proposal. He suggested that the figure 10^{40} is in fact not a constant, but a variable that changes with time and is connected with the age of the universe. According to the evolutionary cosmology, which holds that the universe originated with a "big bang," the universe is now about 5×10^9 years, or 10^{17} seconds, old. Of course, a year or a second is an arbitrary unit, and we would prefer an elementary time interval that can be derived from the basic properties of matter and light. A reasonable one is the length of time required by light to travel a distance equal to the radius of an elementary particle. Since all the particles have radii of about 3×10^{-13} centimeter, and since the velocity of light is 3×10^{10} centimeters per second, this elementary time unit is 3×10^{-13} divided by 3×10^{10} , or 10^{-23} second. To express the age of the universe in this elementary time unit we divide its age in seconds, 10^{17} , by 10^{-23} and obtain the number 10^{40} ! Thus, said Dirac, the large ratio of electric to gravitational forces is characteristic of the present age of the universe. When



SIGN OF SPACE CURVATURE depends on type of gravitational field. Observers on three planets (*left*) would detect positive curvature because of deflection of light rays by the sun. Observers on periphery of merry-go-round (*right*) would detect negative curvature.

the universe was half as old as it is now, this ratio was also half of its present value. Since there are good reasons to assume that the elementary electric charge does not change with time, Dirac concluded that the gravitational constant must be decreasing, and that this decrease may be associated with the expansion of the universe and the steady rarefaction of the material that fills it.

If the gravitational constant really has

been decreasing, or in other words if the force of gravity has been growing weaker, then our solar system must have been expanding along with the universe. In earlier times the earth would have been nearer the sun and therefore hotter than it is now. When Dirac put forward the idea, the solar system was thought to be about three billion years old. Edward Teller, now at the University of California, pointed out that on such a time

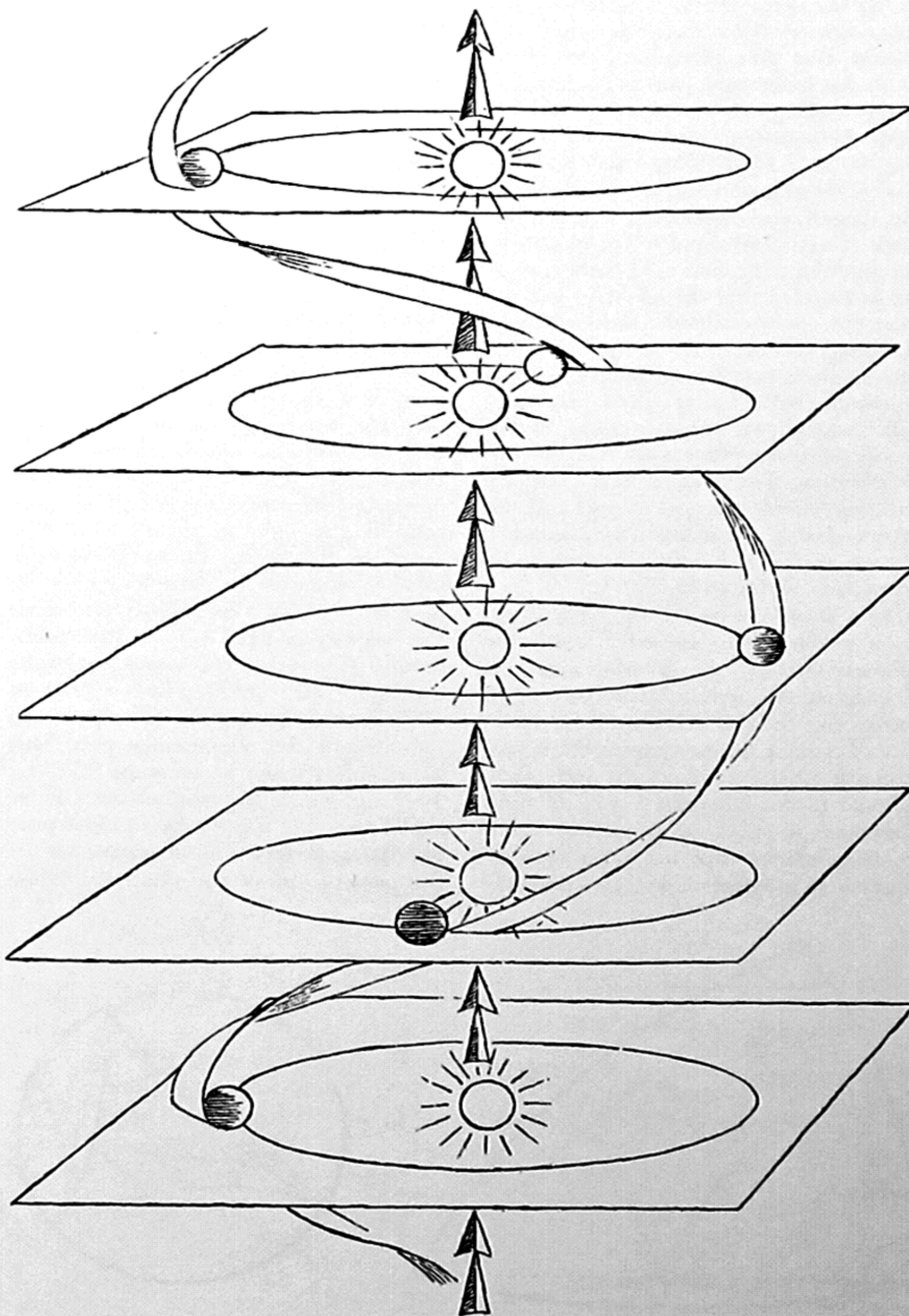
scale the earth would have been 50 degrees hotter than the boiling point of water during the Cambrian era, when well-developed marine life existed. Now it seems that the solar system may be five billion or more years old, in which case the Cambrian oceans, though hot, would not have been vaporized. So the objection loses its force, provided that Cambrian plants and animals could live in very hot water.

Antigravity

In one of his stories H. G. Wells describes a British inventor, Mr. Cavor, who found a material, called cavorite, that was impenetrable to the force of gravity. Just as sheet copper can shield an object against electric forces and sheet iron can shield against magnetism, a sheet of cavorite placed under a material body would shield it from the gravitational pull of the earth. Mr. Cavor built a large gondola surrounded by cavorite shutters. One night when the moon was high, he got into the ship, closed the shutters facing the ground and opened those facing the moon. Cut off from terrestrial gravity and subjected only to the attraction of the moon, the gondola soared into space and eventually deposited Mr. Cavor on the surface of our satellite.

Why is such an invention impossible? Or is it? There is a profound similarity between Newton's law of universal gravity and the laws that govern the interactions of electric charges and magnetic poles. If one can shield electric and magnetic forces, why not gravity? To answer this question we must consider the mechanism of electric and magnetic shielding. Each atom or molecule in any piece of matter is a system of positive and negative electric charges; in conducting metals there are numbers of negative electrons that are free to move through the crystal lattice of positively charged ions. When a metal is placed in an electric field, the free electrons move to one side of the material, giving it a negative charge and leaving the opposite side positive. This polarization produces a new electric field, which is directed opposite to the original field. Thus the two can cancel each other. Similarly, magnetic shielding depends on the fact that the atoms of magnetic materials are tiny magnets, with north and south poles that line up so as to produce a field that opposes an external magnetic field. Here also the shielding effect arises from polarization of atomic particles.

Gravitational polarization, which



FOUR-DIMENSIONAL PATH (*spiral*) of a planet in space-time is a geodesic. It is rendered schematically by showing space in two dimensions and measuring time vertically.

could make possible shielding against the force of gravity, requires that matter be constituted of two kinds of particles: some with positive gravitational mass, which are attracted by the earth, and some with negative gravitational mass, which are repelled. Positive and negative electric charges and north and south magnetic poles are equally abundant in nature, but particles with negative gravitational mass are as yet unknown, at least within the structure of ordinary atoms and molecules. Therefore ordinary matter cannot be gravitationally polarized and cannot act as a gravity shield.

There is, however, another kind of matter—antimatter—that in many ways is the reverse of ordinary matter, including its electric and magnetic properties. Perhaps antiparticles also have negative mass. At first sight this might seem an

easy point to decide. One has only to watch a horizontal beam of antineutrons, say, emerging from an accelerator and see whether the beam bends down or up in the gravitational field of the earth. In practice the experiment cannot be done. The particles produced by accelerators move almost at the speed of light; in a kilometer of horizontal travel gravity would bend them, whether up or down, only about 10^{-12} centimeter, the diameter of an atomic nucleus. Nor can they be slowed down by letting them collide with the nuclei of a "moderator" material, as neutrons are slowed in atomic piles. If antiparticles collide with their ordinary counterparts, both disappear in material annihilation. Thus from the experimental point of view the question as to the sign of the gravitational mass of antiparticles remains painfully open.

From the theoretical point of view it is open too, since we do not have a theory that relates gravitational and electromagnetic interactions. If a future experiment should demonstrate that antiparticles do have a negative gravitational mass, it will deliver a mortal blow to the entire relativistic theory of gravity by disproving the principle of equivalence. An antiapple might fall up in a true gravitational field, but it could hardly do so in Einstein's accelerated spaceship. If it did, an outside observer would see it moving at twice the acceleration of the ship, with no force at all acting on it. The discovery of antigravity would therefore force upon us a choice between Newton's law of inertia and Einstein's equivalence principle. The author earnestly hopes that this will not come to pass.

The Author

GEORGE GAMOW is professor of physics at the University of Colorado. After receiving his doctoral degree in nuclear physics from the Leningrad State University in 1928, Gamow continued his studies under Niels Bohr at the University of Copenhagen and later under Ernest Rutherford at the University of Cambridge. He came to the U. S. in 1934 and taught physics at George Washington University until 1956, when he went to the University of Colorado. A prolific popularizer of science, Gamow was awarded the Kalinga Prize in 1956 for his interpretation of science for the layman. He has published a dozen books in

23 languages, with three more going to press. This is his 11th article for **SCIENTIFIC AMERICAN**.

Bibliography

SIR ISAAC NEWTON'S MATHEMATICAL PRINCIPLES OF NATURAL PHILOSOPHY AND HIS SYSTEM OF THE WORLD. Edited by Florian Cajori. University of California Press, 1947.

DIALOGUES CONCERNING TWO NEW SYSTEMS. Galileo Galilei. Dover Publications, Inc., 1953.

MEANING OF RELATIVITY. Albert Einstein. Princeton University Press, 1956.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound **SCIENTIFIC AMERICAN** Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

DATE LOANED

Acc. No. _____

This book may be kept for 14 days. An over - due charge will be levied at the rate of 10 Paise for each day the book is kept over - time.

[illegible]

OPTICAL MASERS

by Arthur L. Schawlow

These devices generate light in such a manner as to open up a whole new realm of applications for electromagnetic radiation. The salient feature of the light they produce is that its waves are all in step.

For at least half a century communications engineers have dreamed of having a device that would generate light waves as efficiently and precisely as radio waves can be generated. The contrast in purity between the electromagnetic waves emitted by an ordinary incandescent lamp and those emitted by a radio-wave generator could scarcely be greater. Radio waves from an electromagnetic oscillator are confined to a fairly narrow region of the electromagnetic spectrum and are so free from "noise" that they can be used for carrying signals. In contrast, all conventional light sources are essentially noise generators that are unsuited for anything more than the crudest signaling purposes. It is only within the last year, with the advent of the optical maser, that it has been possible to attain precise control of the generation of light waves.

Although optical masers are still very new, they have already provided enormously intense and sharply directed beams of light. These beams are much more monochromatic than those from other light sources; at their best optical masers rival the very finest electronic oscillators as a source of a single frequency. The development of optical masers is moving so rapidly that they should soon be ready for a wide variety of applications. These may range from space communications and radar to accelerating specific reactions in chemical technology.

To appreciate the limitations of light waves as they are ordinarily found, let us consider how they are produced. All light sources—incandescent lamps, arcs and so on—are essentially hot matter. In the familiar neon tube, it is true, the glass walls remain cool, but the electrons and gas atoms within the tube are accelerated to the high speeds normally

associated with high temperatures. The atoms are continuously "pumped" to an excited state; then they fall back, losing energy and radiating visible light. They fall back, however, one atom at a time. The disorderly atomic motion we associate with a heated gas is paralleled by a disorderly outpouring of light quanta, or photons. Just which atoms radiate at any instant is purely a random affair. The excited electrons in the hot tungsten filament of an incandescent lamp also radiate randomly and independently.

Thus the light that comes from any conventional light source is called spatially incoherent. This means that the light emerges in a jumble of tiny, separate waves that reinforce or cancel each other in random fashion; the wave front so produced varies from point to point and changes from instant to instant. The wave front resembles one that would be produced by throwing a handful of pebbles into a pool. If, on the other hand, only a single pebble were dropped into the pool, a coherent circular wave front would be produced. By the same token one can imagine a point source of light that could generate coherent waves whose fronts would form spherical surfaces. Alternatively, a suitable source might generate coherent light waves whose fronts were plane surfaces; at every point on the plane the strength of the electric field would be the same. As the wave fronts traveled past a given point in space, the field strength would be seen to rise and fall smoothly and rhythmically in phase, swinging from positive to negative in value.

If a conventional electronic oscillator that produces radio waves is connected to a small radiator of suitable design, the radiator will send out spherical coherent waves. If one wishes, the oscillator can be connected to feed a number of radiat-

ing antennas that will send out a directional wave, much like a plane wave.

To obtain a directional wave from an incoherent light source one must start with a source of small dimensions. Then, by placing a screen with a hole in it some distance from the source, we can select the segment of the wave that happens to be going in the desired direction. Alternatively, the light emitted by a small source can be focused with a larger mirror or lens to yield a beam with sides that are roughly parallel. The sides of a beam produced by an arc lamp and a six-foot mirror diverge at an angle of about one degree. As we shall see, the output of an optical maser is both more directional and more coherent.

Perhaps the most important limitation of ordinary light sources is their inherent low brightness. No matter how high their temperature, they cannot emit more energy than a perfect radiator. The theoretical output of a perfect radiator, called a black body, is given by the famous black-body radiation curve first derived by Max Planck. The visible surface of the sun, for example, behaves much like a black body with a temperature of 6,000 degrees centigrade. The sun's total radiation, at all wavelengths, is seven kilowatts per square centimeter of its surface, and no matter how we collect and concentrate sunlight it is impossible to achieve any greater radiation density.

Although seven kilowatts may seem a substantial amount of energy, it is really not very much if one considers the tremendous width of the solar spectrum. To bring this point home let us compare the width of the visible portion of the electromagnetic spectrum with the width of a standard television channel, which is four megacycles. A little calculation

shows that the visible region between the wavelengths of 4,000 and 7,000 angstrom units could contain 80 million television channels. In other words, each angstrom unit is about 100,000 megacycles wide. If one were able to filter out a narrow band of green light one megacycle wide from the region where the sun has its peak energy output (4,800 angstrom units), one would find that each square centimeter of solar surface produces only .00001 watt. To obtain as much as a single watt of green light one megacycle wide one would have to collect and filter the output from 10 square yards of solar surface. In contrast, man-made transmitters operating in the television region of the radio spectrum can easily generate 10,000 watts in a band much less than one megacycle wide.

Ordinary light sources are like the sun: they are broad-band noise generators that spread their output over a wide range of frequencies without supplying much power at any particular frequency. Even gas-discharge lamps, which emit light at a restricted number of narrow lines in the spectrum, do not approach the best electronic oscillators as sources of power at a single frequency.

There has been, of course, a great effort to extend electronic oscillators to shorter wavelengths. The length of the shortest waves that can be generated by electronic means is about one millimeter, or 10 million angstrom units. Any attempt to reach shorter wavelengths with conventional electronic designs meets with formidable difficulties. Foremost among these is the difficulty of fabricating the resonant structures that tune the oscillator. These structures can seldom be much larger than a wavelength in size. At millimeter wavelengths they are already so small that they are hard to make with uniform accuracy. To produce optical wavelengths, which are three orders of magnitude shorter, a radically different approach is needed.

An attractive solution to the problem would be to stop trying to build these tiny resonators and to replace them by atomic or molecular resonators. Nature has provided us with a wide variety of such resonators through the entire infrared, visible and ultraviolet spectrum. Indeed, engineers are accustomed to using atomic oscillators in gas-discharge lamps. A single atom, however, radiates very little power, and that only intermittently. What is needed is some way to synchronize a large number of atoms so that they can work together to produce a powerful, coherent wave.

Such an approach has been made pos-

sible by the maser principle, discovered by Charles H. Townes at Columbia University [see "The Maser," by James P. Gordon; SCIENTIFIC AMERICAN Offprint 215]. Maser stands for "microwave amplification by stimulated emission of radiation." The original maser, completed in 1954 by James P. Gordon, H. J. Zeiger and Townes, used the vibrations of ammonia molecules to provide microwave oscillations of precisely determined frequency. Subsequently Nicolaas Bloembergen of Harvard University indicated a practical way to build the so-called three-level solid-state maser for use as a low-noise microwave amplifier. The first maser of this type was built at the Bell Telephone Laboratories by George Feher, H. E. D. Scovil and H. Seidel, and since then many others have been constructed. Radio astronomers have found them extremely valuable for amplifying very faint radio signals from space. Last year masers were also used to amplify the weak signals received on the bounce from the *Echo* satellite.

Stimulated emission, which is the basis of maser operation, is the reverse of the process in which electromagnetic waves, or photons, are absorbed by atomic systems. When a photon is absorbed by an atom, the energy of the photon is converted to internal energy of the atom. The atom is then raised to an "excited" quantum state. Later it may radiate this energy spontaneously, emitting a photon and reverting to the "ground" state or to some state in between. During the period in which the atom is still excited it can be stimulated to emit a photon if it is struck by an outside photon having precisely the energy of the one that would otherwise be emitted spontaneously. As a result the incoming photon, or wave, is augmented by the one given up by the excited atom. More important and more remarkable, the wave, upon release, falls precisely in phase with the wave that triggered its release. This phenomenon lies at the heart of the maser principle [see illustration on following page].

The problem in designing a maser is to prepare an "active medium" in which most of the atoms can be placed in an excited state, so that an electromagnetic wave of the right frequency passing through them will stimulate a cascade of photons. There must be an excess of excited atoms to enable stimulated emission to predominate over absorption. Atoms are raised to an excited state by injecting into the system electromagnetic energy at a wavelength different from

the stimulating wavelength; the activating process is called "pumping."

Once an active medium has been prepared it can be enclosed in a reflecting box, or cavity resonator. Then a wave that starts out at one wall of the box will grow in amplitude until it reaches another wall, where it will be reflected back into the mass of excited atoms. Inevitably there are losses at the walls due to imperfect reflection. If the amplification by stimulated emission is great enough to make up for these reflection losses, a steady wave will build up in the box. At centimeter wavelengths it is not difficult to build a box having the dimensions of a wavelength and so designed that the wave will build up with only one mode of oscillation. A single mode of oscillation corresponds to a single frequency of output; extraneous modes create extra frequencies, or noise, and compete with the desired mode in extracting energy from the supply of excited atoms.

At optical wavelengths a single-wavelength resonator would have dimensions inconveniently small. To surmount this problem, Townes and the author proposed in 1958 that a maser for optical wavelengths could be built by making a special kind of resonator with dimensions thousands of times greater than the emission wavelength but which nevertheless would favor a particular mode of oscillation. In the optical maser the reflecting box is replaced by a device or structure with two small mirrors facing each other. A wave that starts out near one mirror and travels along the axis of the system will grow by stimulated emission until it reaches the other mirror. There it will be reflected back into the active medium so that growth can continue. If the gain on repeated passages is enough to make up for the losses at the mirrors, a steady wave will be built up. If one of the mirrors is semi-transparent, a portion of the wave can escape through it, constituting the output of the maser.

It is perhaps evident that a wave inclined at an angle to the axis will leave the system after only a few reflections, or perhaps without ever striking one of the mirrors. Such a wave will not have the same opportunity to build up as does a wave that travels straight along the axis of the system. Like other maser oscillators, the optical maser Townes and I described would be triggered off by the first photons to be emitted spontaneously after the system had been "pumped up" to the active state. (A maser designed to operate as an amplifier, on the

other hand, uses an input signal as a stimulating wave.)

We had every reason to expect that the output of an ideal optical maser constructed in this manner would be extremely directional, very powerful, essentially monochromatic and, above all, coherent. The output would be directional because the only waves emitted would have had to make repeated passages—perhaps thousands—without deviating very far from the axis of the maser. It would be very powerful because stimulated emission forces excited atoms to radiate much earlier than they would spontaneously. It would be very monochromatic because stimulated emission is a resonant process and takes place most strongly at the center of the band of frequencies that can be emitted in spontaneous radiation. These favored frequencies will in turn cause emission at the same frequency, so that the wave built up in the maser will contain only an extremely narrow range of frequencies or wavelengths.

Finally, the output of the optical maser, if it is a good approximation of

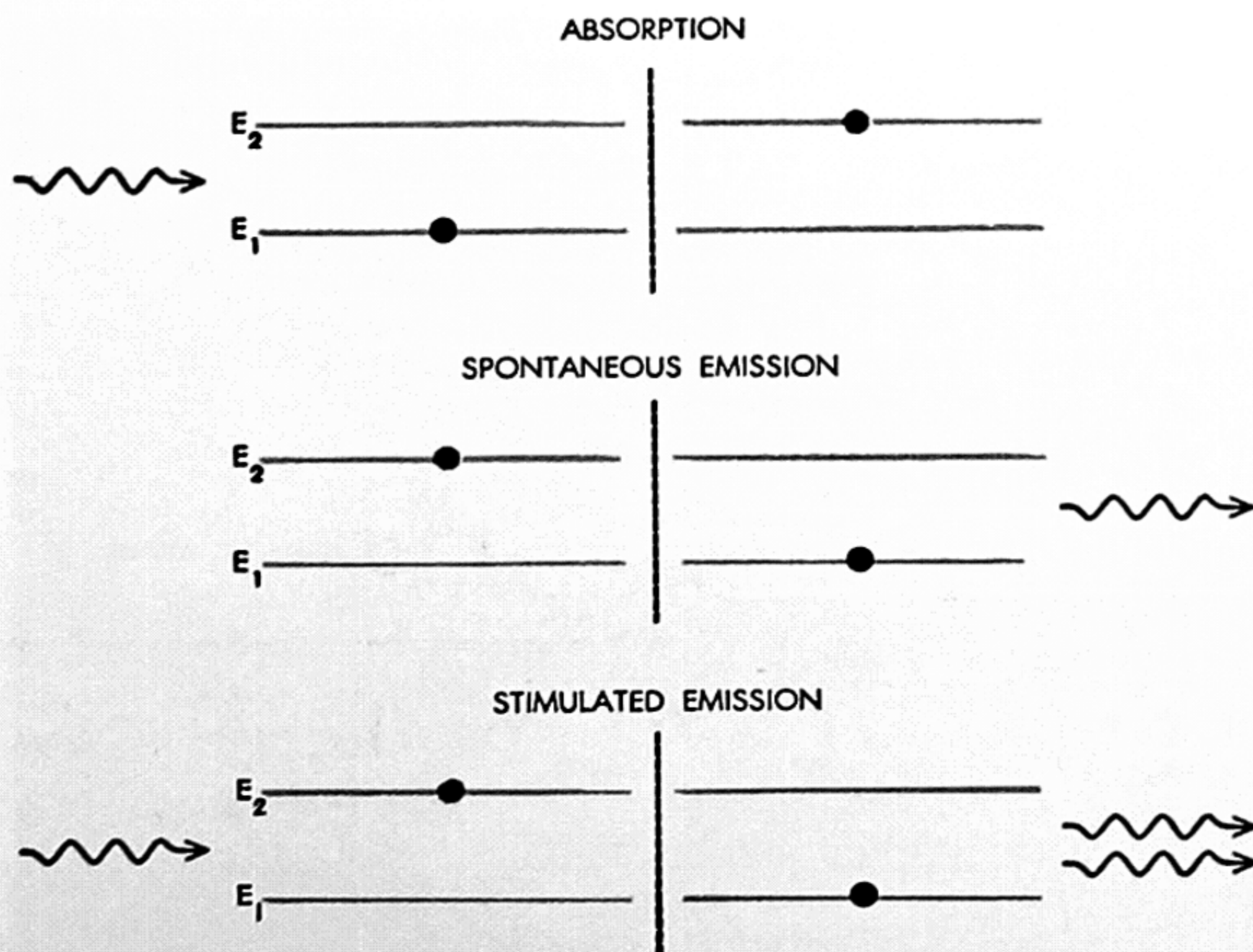
a plane wave traveling in a single direction, will be spatially coherent because all the wave fronts are planes perpendicular to the direction of propagation. Since the maser output is nearly monochromatic, it also has time coherence. This means that there is a fixed phase relation between the portion of the wave emitted at one instant and the wave emitted after a fixed time interval. For a wave whose period is one second, crests follow each other at one-second intervals. On the other hand, if the frequency varies, the interval between crests is irregular. The more nearly the wave holds to a single, fixed frequency, the more nearly it exhibits time coherence.

Testing these predictions required preparation of an active medium that would actually display maser action in the optical region of the spectrum. The first announcement of success was made last July by T. H. Maiman of the Hughes Aircraft Company, whose device used a ruby crystal. Between July and the end of 1960 four other substances were tested successfully by various investiga-

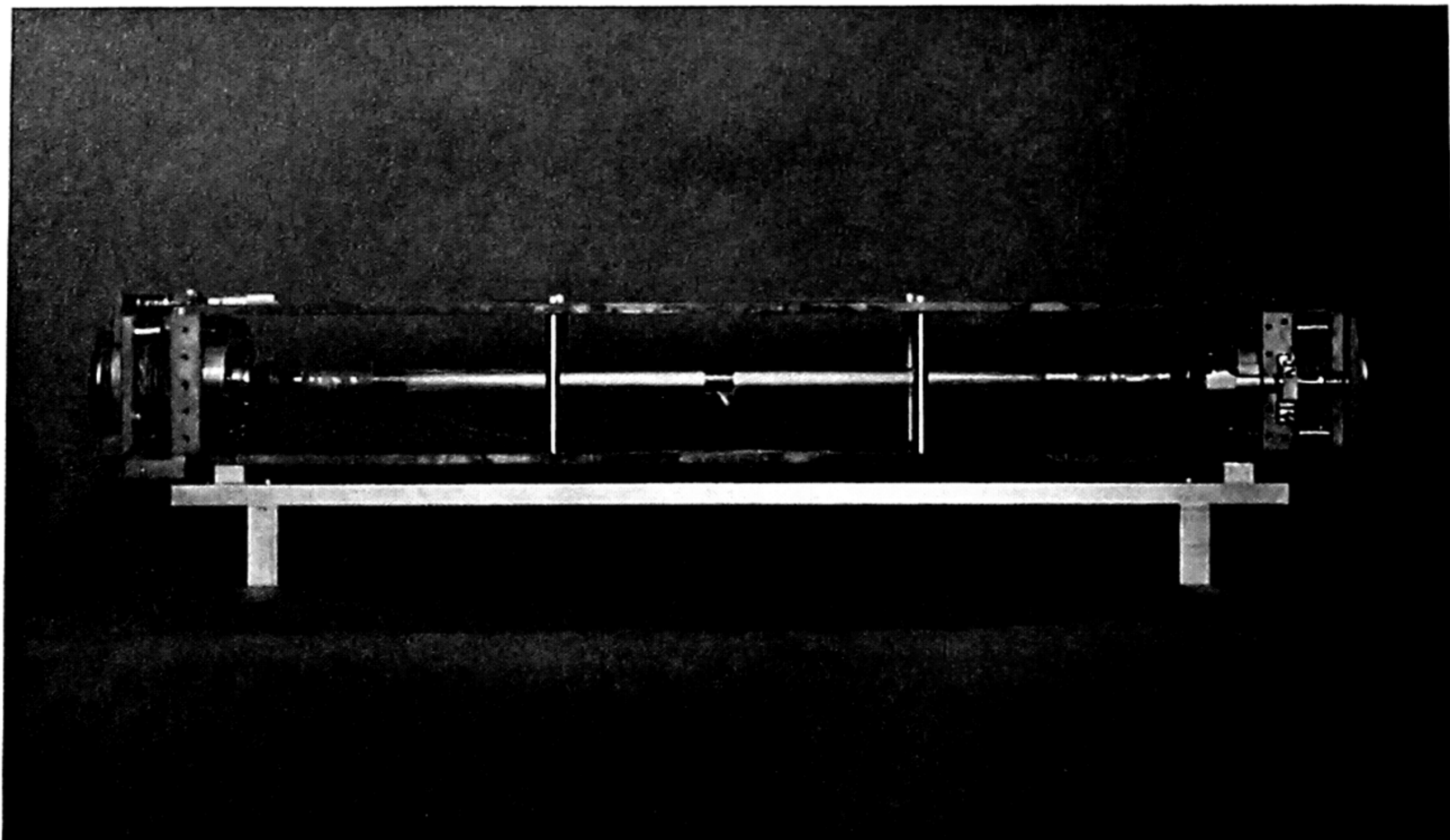
tors. These devices all embodied the concept of reflecting end walls, described above. At last count optical maser oscillations had been obtained at 11 different wavelengths. It seems likely that these wavelengths will soon be joined by many others.

Maiman's ruby maser is typical of those using crystals. Ruby is aluminum oxide in which a few of the aluminum atoms have been replaced by chromium atoms; the more chromium, the deeper the color. Maiman used a pale pink ruby containing about .05 per cent chromium. The color results from the fact that the chromium atoms in the crystal absorb a broad band of green and yellow light, along with ultraviolet light, and let only the red and blue pass through. Moreover, the light that is absorbed raises the chromium atoms to an excited state from which two steps are required to carry them back to the ground state. In the first step they give up some of their energy to the crystal lattice and land temporarily in what is called a metastable state. If they are not subjected to stimulation, their stay at this level lasts a few milliseconds while they drop at random to the ground state. Photons emitted during this final drop have a wavelength (at room temperature) of 6,943 angstrom units, which accounts for the characteristic red fluorescence of ruby crystals. In an optical maser, however, the first few photons released at this wavelength stimulate the still excited chromium atoms to give up photons and tumble to the ground state much sooner than they would normally; the result is a cascade of photons all at the 6,943-angstrom wavelength.

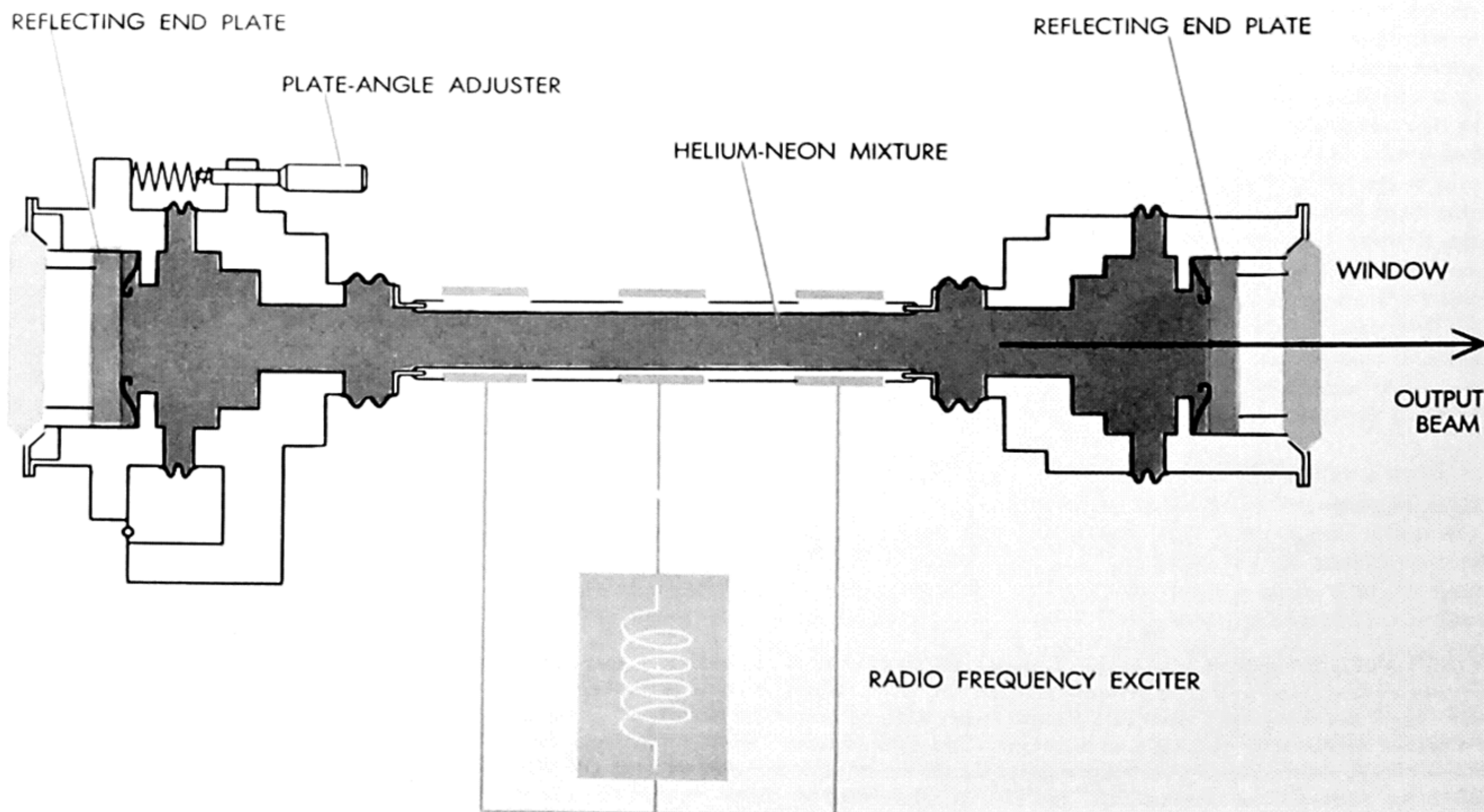
For use in an optical maser the pink ruby is machined into a rod about four centimeters long and half a centimeter across. Its ends are polished optically flat and parallel and are partially silvered. The rod is placed near an electronic flash tube that provides broadband pumping light. What Maiman discovered before anyone else is that the most powerful of these lamps, when connected to a large power supply, is capable of raising most of the chromium atoms to the excited state. Up to a certain critical flash intensity, all that happens is that the ruby emits a burst of its typical red fluorescence spread over the usual decay period for the excited atoms. But above a critical level maser action takes over, and an intense red beam, lasting for about half a thousandth of a second, flashes out from the partially silvered ends of the rod. This shows that a



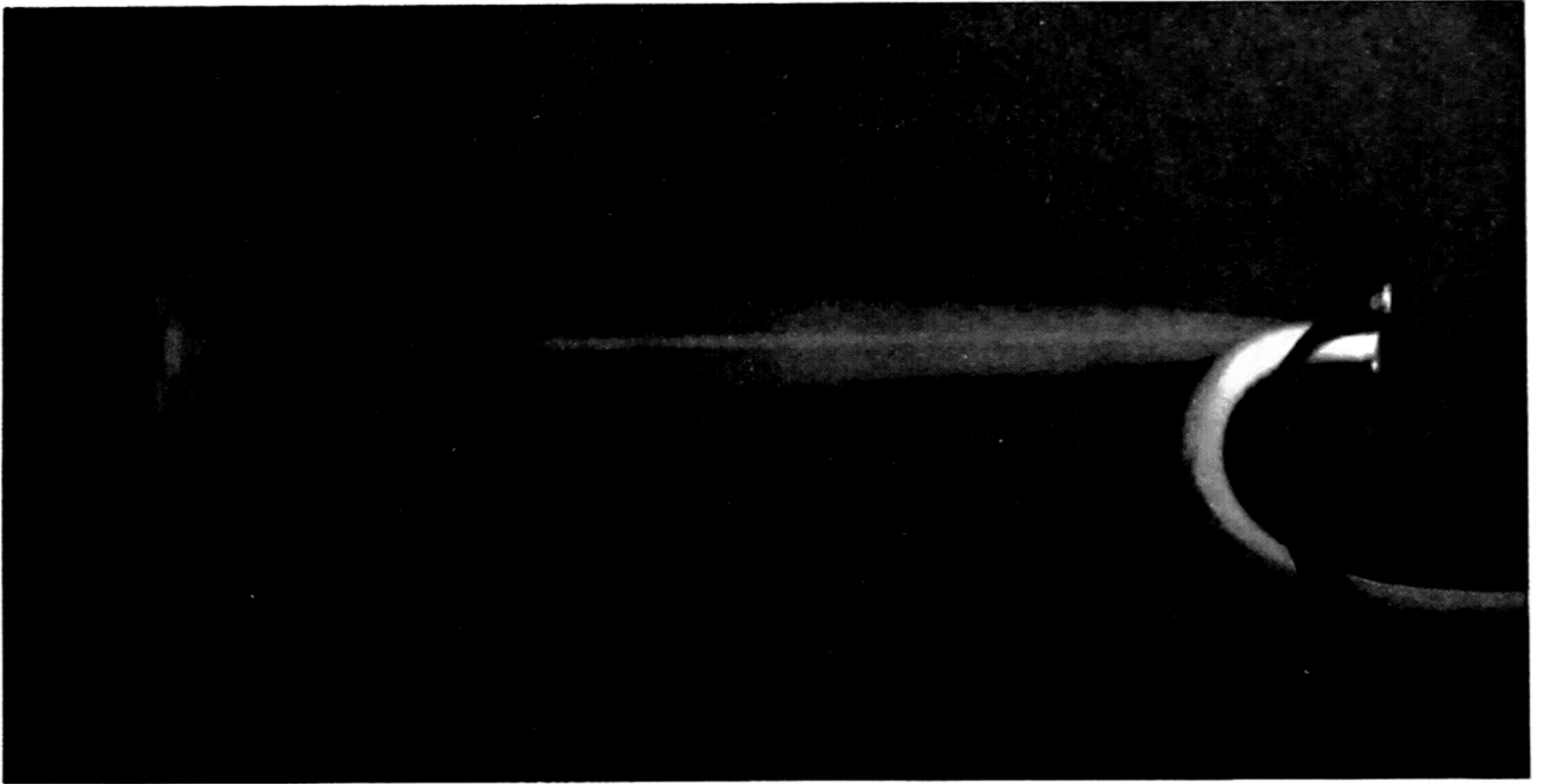
STIMULATED EMISSION of photons (*bottom*), the basis of maser operation, is contrasted with absorption (*top*) and spontaneous emission (*middle*). When an atom in the "ground" state (*black dot at top left*) absorbs a photon (*wavy colored arrow*), it is excited, or raised to a higher energy state (*gray dot at top right*). The excited atom (*middle left*) may then radiate energy spontaneously, emitting a photon and reverting to the ground state (*middle right*). An excited atom (*bottom left*) can also be stimulated to emit a photon when it is struck by an outside photon. Thus in addition to the stimulating photon there is now a second photon of the same wavelength (*bottom right*) and the atom reverts to the ground state.



GAS OPTICAL MASER employs a helium-neon mixture to produce an infrared beam. Reddish glow results from gas discharge of mixture. The gas maser, designed by Ali Javan, W. R. Bennett, Jr., and D. R. Herriott, was the first to operate continuously.



OPERATION OF GAS MASER depends on right mixture of helium and neon gases to provide an active medium. Radio frequency exciter puts energy into the medium. The output beam is built up by repeated passes back and forth between reflecting end plates.



BEAM FROM RUBY OPTICAL MASER makes a thin streak of red light as it passes through smoke. Upper end of curved cooling

tube at right is attached to front end of the maser housing, which here cannot be seen. The ruby crystal is mounted in the housing.



BRILLIANT BLUE-WHITE FLAME at left is incandescent carbon vapor produced by focusing a ruby-maser beam on a carbon target. The beam heated a spot on the target to about 8,000 degrees centi-

grade in .0005 second. The large curved object to the right is the same cooling tube seen in the photograph at top. The small lens used to focus the beam is mounted at the left of the tube.

sufficient excess of atoms has been pumped up to the excited state to make up for the losses at the ends.

In 1959 the author had predicted that it should be possible to build an optical maser using a dark red ruby that contains about 10 times as much chromium as the pink ruby. It was predicted that at this higher concentration maser action would take place simultaneously at two different wavelengths—7,009 and 7,041 angstroms [see illustration on opposite page]. This mode of operation was recently demonstrated in a maser built by G. E. Devlin and the author, and in another built by I. Wieder and L. R. Sarles of Varian Associates. Still other solid optical masers, using samarium or uranium ions in a calcium fluoride crystal, have been constructed by P. P. Sorokin and M. J. Stevenson of the International Business Machines Research Laboratory. These masers oscillate at wavelengths of 7,080 and 25,000 angstroms respectively.

All these masers were first operated in short bursts, but they seem potentially capable of continuous operation. The active medium used by Maiman, however, is less well suited than the others in which the stimulated emission comes in a transition to an intermediate energy level that is some distance above the ground state. It is not necessary, there-

fore, to expend a lot of energy pumping half the atoms out of the ground state so that emission can predominate over absorption. In the newer materials the intermediate state—where the atoms land after emitting photons of the desired frequency—can be emptied simply by cooling. Hence the active medium contains very few atoms “tuned” to absorb the photons produced by maser action. Only enough pumping is needed for this action to begin.

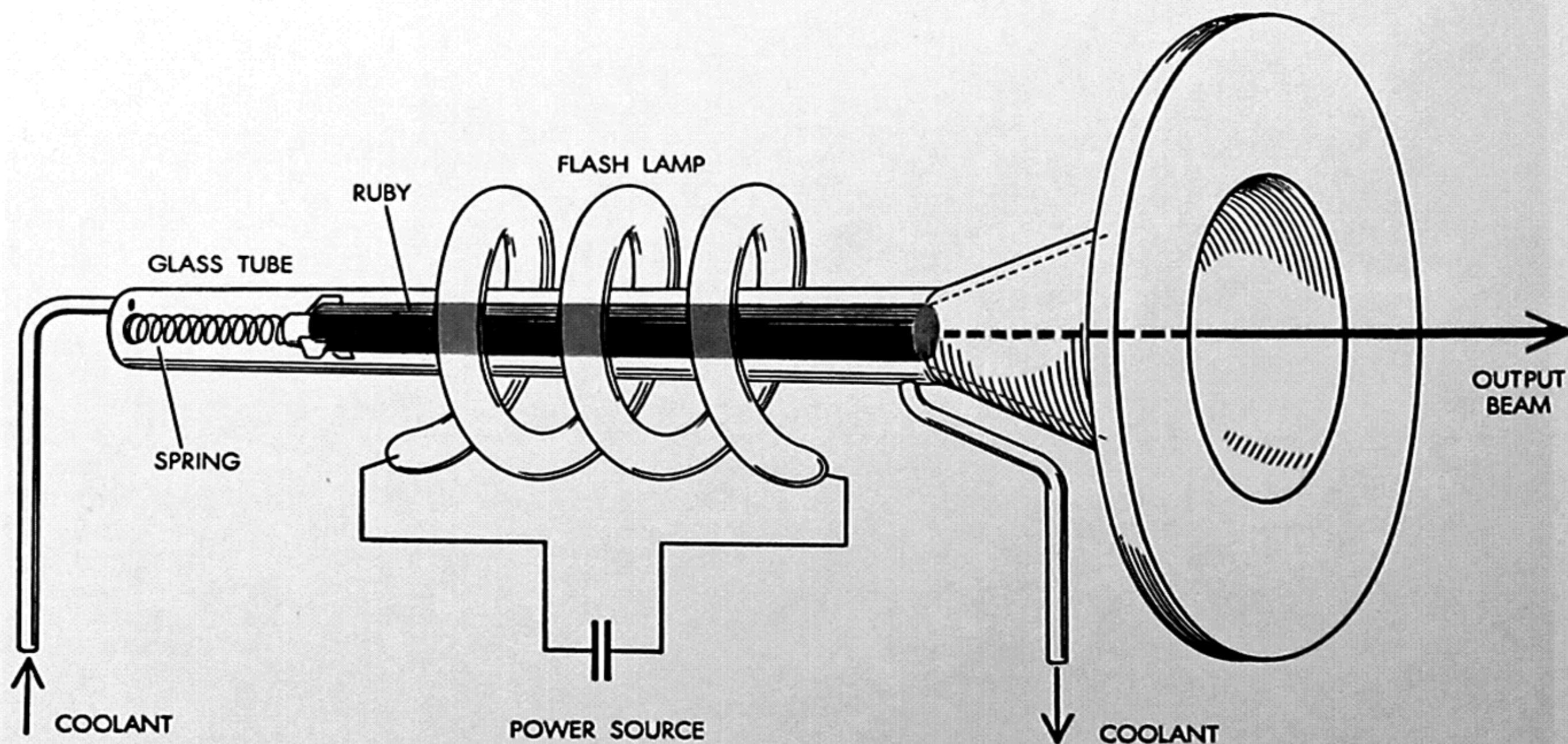
A totally different way to obtain excited atoms for an optical maser employs gas atoms in an electric glow-discharge under very special conditions. In a mixture of helium and neon gas it is possible to secure maser oscillation at several wavelengths in the infrared region around 10,000 angstroms. This system was proposed in 1959 by Ali Javan of the Bell Telephone Laboratories. A successful prototype, constructed by him in collaboration with W. R. Bennett, Jr., and D. R. Herriott, was demonstrated early this year. The principal feature of this maser is its ability to operate continuously and with a very low energy input—about 50 watts in the first model.

In Javan's maser the stimulated emission occurs when a neon atom falls between two intermediate levels, the lower of which is well above the ground state.

Only a modest input of energy is required to produce a gas discharge—essentially like that in an ordinary neon sign—and this in turn provides a continuous supply of neon atoms at the proper level of excitation needed to produce a continuous maser beam [see illustrations on page 622]. As in the ruby masers, the beam is built up and made coherent by bouncing back and forth between reflecting end plates.

The helium-neon maser exemplifies the increasing subtlety of maser designs. The energy needed to raise the neon atom to an excited state is not supplied directly by an incoming photon; it is supplied by collision with an excited helium atom. Many other possibilities remain to be explored. Energy levels suitable for masers may be found in many different kinds of system. For example, in the infrared region spectral lines are produced by vibrations of gas molecules, by vibrations of crystals and by electronic excitations of certain atoms in crystals. Which of these may be usable in a maser can be discovered only by a detailed study of a system's spectrum.

Now that optical masers have been built, how closely do they match expectations in power, directionality, coherence and the narrowness of the band of wavelengths produced? We know



RUBY MASER is powered by a flash lamp, which provides pumping energy. Output beam is emitted through partially silvered end of ruby crystal; other end is completely silvered. Beam builds up

by repeated reflection between the ends. Liquid nitrogen is used to cool the ruby, though it can also be operated at room temperature. Only the front end of the maser housing (right) is shown.

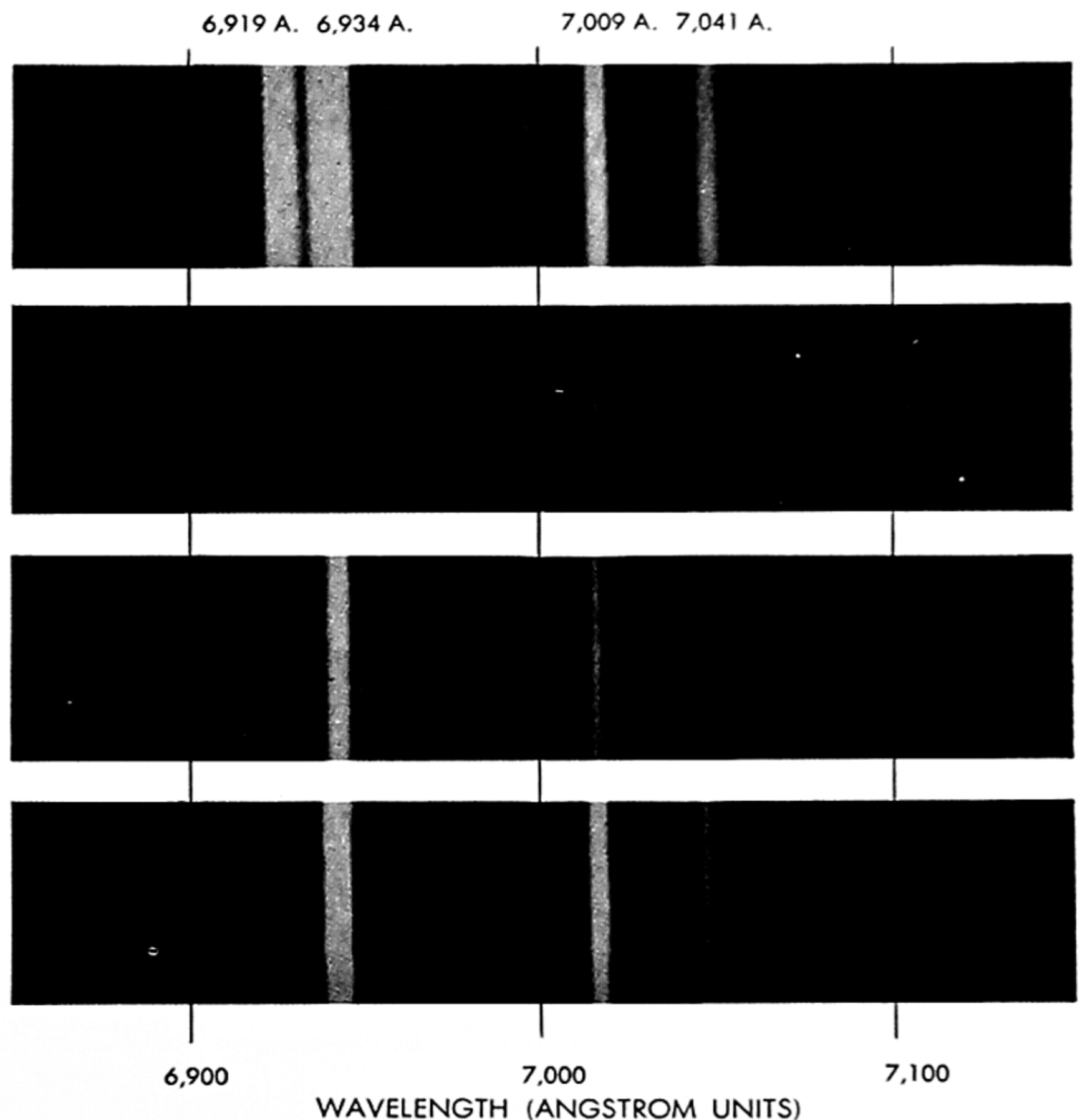
most about the pink ruby. In short bursts its power output reaches 10,000 watts for a beam measuring less than a square centimeter in cross section. The sides of the beam are parallel to within less than half a degree; at lower power the divergence drops to about a twentieth of a degree. The latter divergence corresponds to a spread of only five feet per mile, and it could be reduced by running the beam through a telescope backward. With telescopic demagnification it should not be difficult to project on the moon a spot of light only two miles in diameter.

If desired, the maser's power can be focused to produce intense heating. For instance, a lens with a focal length of one centimeter will focus the beam to a spot only a hundredth of a centimeter in diameter, corresponding to an area of one ten-thousandth of a square centimeter. In this spot the maser beam will deliver power at a density of 100 million watts per square centimeter. Brief though the flash is, its power is thousands of times greater than could be obtained by focusing sunlight and is enough to melt or vaporize a spot on the surface of even the most refractory material. This was first demonstrated by my colleague W. S. Boyle [see bottom illustration on page 625].

It is not surprising that the ruby maser falls short of ideal performance in some respects, particularly in the narrowness of the band of wavelengths it produces. Because it is violently pulsed, the ruby rod heats appreciably. Nevertheless, when the maser threshold level is reached, the band of wavelengths narrows to about .02 angstrom, or about 1,000 megacycles. This is as narrow as the sharpest spectral line from any nonmaser light source.

The ruby is far surpassed in the narrowness of its output of wavelengths by the gas maser of Javan, Bennett and Herriott. This maser produces spectral lines less than one kilocycle wide at a carrier frequency of 100,000 megacycles. The gas maser's power output per kilocycle of band-width is about 100 million times that of a square centimeter of the sun's surface. It is possible that the frequency of a maser's output will drift slightly over a period of time, but over a short period it is remarkably stable; in the radio range this stability is equaled only by the finest frequency standards and by atomic clocks.

Of all the properties of the optical maser, none is more striking than the spatial coherence of its light. This is readily demonstrated by using the maser



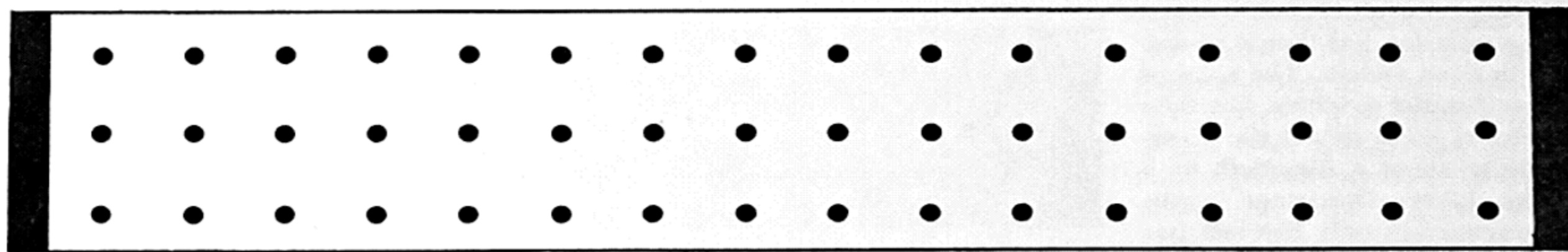
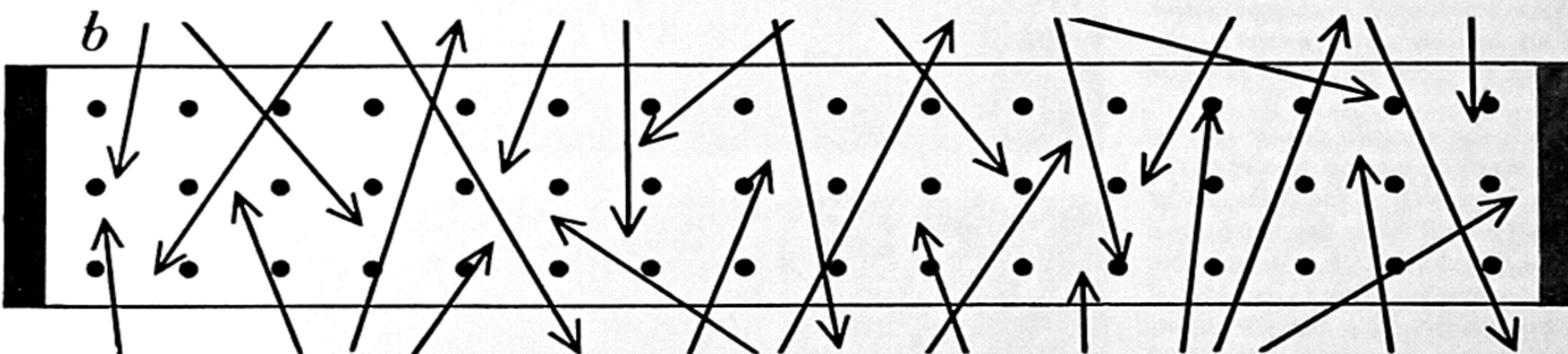
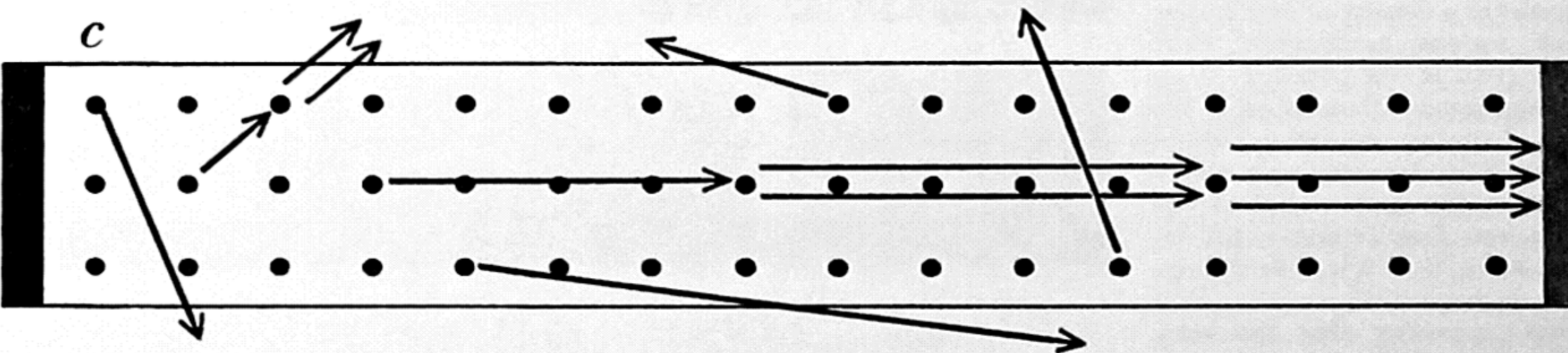
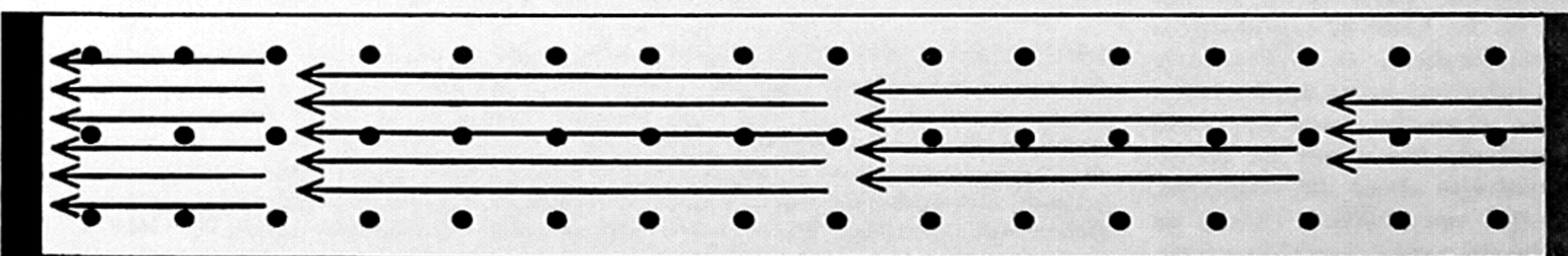
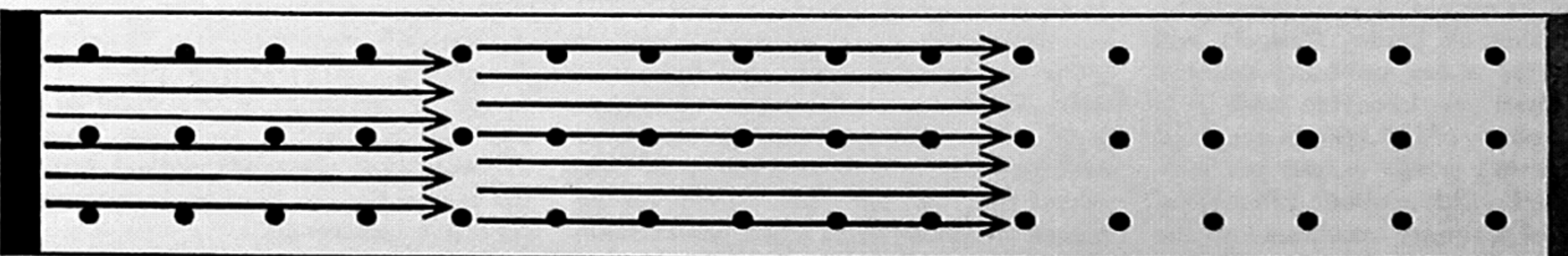
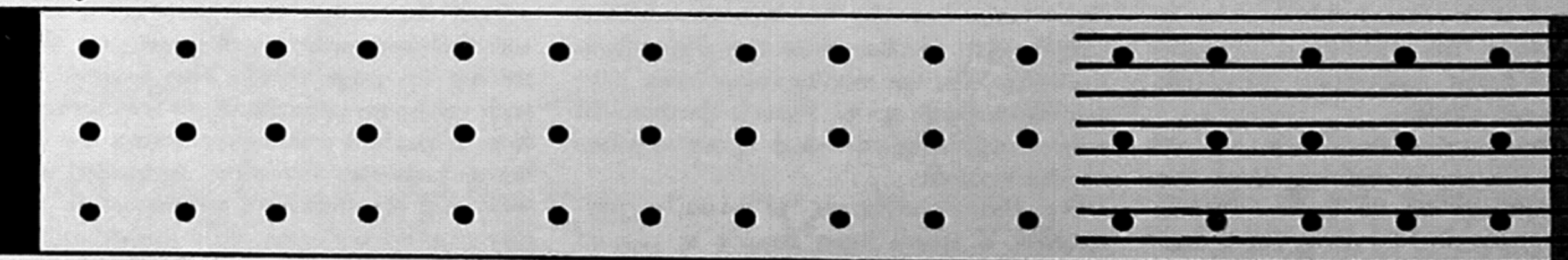
RUBY-MASER SPECTRA in three lower photographs are compared with spectrum of spontaneous (unstimulated) ruby fluorescence at top. As pumping power reaches first threshold for maser action (*second from top*), the ruby "masers" at 7,009 angstrom units (A.), at two wavelengths as power increases (*third from top*), then at three (*bottom*). Maser oscillation never occurs at 6,919 A. Sequence in which spectral lines appear varies with the maser crystal used and operating conditions. A 30-second exposure was required to photograph the fluorescent spectrum. Three lower photographs show single flashes of .0005 second.

to repeat the classic two-slit interference experiment first performed in 1806 by Thomas Young to show that light consists of waves. In Young's experiment light passes through two parallel slits and then falls on a distant screen. If light waves from one slit reach a point on the screen in phase with light waves from the other slit, the two sets of waves reinforce, producing a bright spot. At a nearby point on the screen, where the light from one slit has traveled half a wavelength farther than the light from the other slit, the waves cancel out, producing a dark spot. Thus a pattern of alternating light and dark spots appears on the screen.

As the experiment is usually performed, a small light source is placed

some distance from the slits, so that the wave fronts reaching them travel nearly perpendicularly to their plane. If the source is too large or too close to the slits, the pattern fails to appear. Young's experiment is therefore a good test for the perpendicularity of wave fronts and for wave coherence.

When the experiment is performed with an optical maser, the slits can be placed directly against the surface from which the beam emerges and a clear interference pattern will result [see illustration on page 629]. The pattern fits well with one calculated on the assumption of perfect coherence across the distance between the slits. Actually, in a ruby rod the region of coherence is usually limited by crystalline imperfections

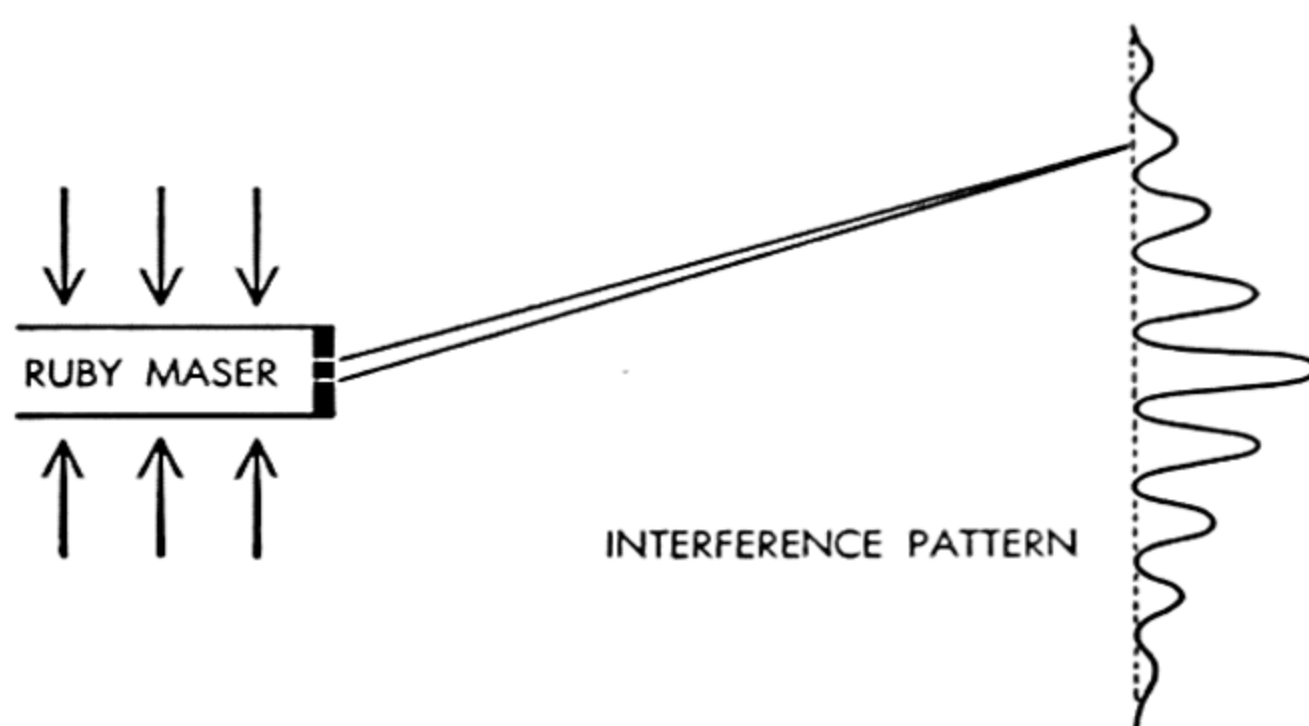
a*b**c**d**e**f*

to about a tenth of the rod's diameter. In the gas maser, however, coherence extends across the whole width of the end plates.

The optical maser is such a radically new kind of light source that it taxes the imagination to canvass its possible applications. Message-carrying, of course, is the most obvious use and the one that is receiving the most technological attention. Signaling with light, although it has been used by men since ancient times, has been limited by the weakness and noisiness of available light sources. The amount of information light signals can carry is thereby quite limited. An ordinary light beam can be compared to a pure, smooth carrier wave that has already been modulated with noise by short bursts of light randomly emitted by the individual atoms in the light source. The maser, on the other hand, can provide an almost ideally smooth wave, carrying nothing but what one puts on it.

If suitable methods of modulation can be found, coherent light waves should be able to carry an enormous volume of information. This is so because the frequency of light is so high that even a very narrow band of the visible spectrum includes an enormous number of cycles per second; the amount of information that can be transmitted is directly proportional to the number of cycles per second and therefore to the width of the band. One must distinguish here between the spectral band-width of the unmodulated maser beam, or carrier wave (which, as we have seen, is extremely narrow), and the band-width after a signal has been impressed upon it.

PHOTON CASCADE in a solid optical maser amplifies a light wave by stimulated emission. Before the cascade begins (a), the atoms in the maser crystal are in the ground state (black dots). Pumping light (black arrows in "b") raises most of the atoms to the excited state (gray dots). The cascade begins (c) when an excited atom spontaneously emits a photon (colored arrow) parallel to the axis of the crystal. (Photons emitted in other directions pass out of the crystal.) The photon stimulates another atom to contribute a second photon. This process continues ("d" and "e") as the photons are reflected back and forth between the ends of the crystal. When the amplification is great enough, some of the beam passes out through partially silvered end of crystal (f).



TWO-SLIT INTERFERENCE EXPERIMENT demonstrates that the light waves leaving a ruby maser are spatially coherent, or in step. When two coherent waves strike a screen, after traveling paths of slightly different length, they reinforce and cancel each other in symmetrical fashion to produce an interference pattern. The photograph of the interference pattern at right was made by D. F. Nelson and R. J. Collins of Bell Telephone Laboratories.

In television transmission the carrier wave (which is also narrow) carries a signal that produces an effective band-width of four megacycles. A single maser beam might reasonably carry a signal with a frequency, or band-width, of 100,000 megacycles, assuming a way could be found to generate such a signal. A signal of this frequency could carry as much information as all the radio-communication channels now in existence. It must be admitted that no light beam will penetrate fog, rain or snow very well. Therefore to be useful in earthbound communication systems light beams will have to be enclosed in pipes.

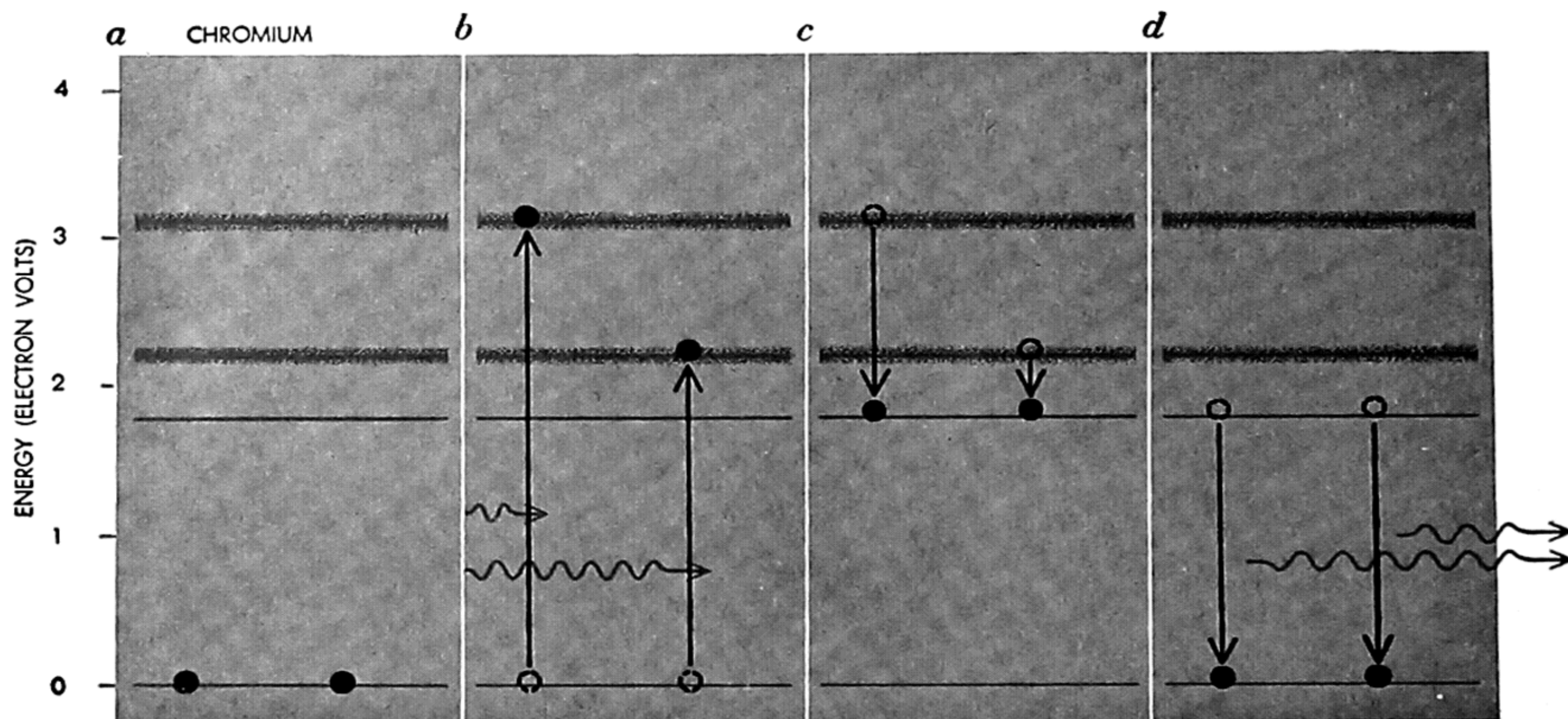
There will certainly be other uses for optical masers. The very intense heat spot produced by focusing an optical maser might be used for fabricating all sorts of electronic devices. For instance, it would be possible to weld a small joint after the joint had been sealed inside a glass envelope. But, in addition to sheer power, the maser provides an intense source of coherent radiation with very high electric-field strength. In such strong fields atoms or molecules may react in strange and unpredictable ways. The beams should therefore be useful in many areas of research. One can also conceive of using maser beams in harmonic generators or mixers. With a suitable mixer one could put in two light waves of different frequency and take out a third wave whose frequency would be the difference of the two original frequencies. In this way it should be possible to synthesize wavelengths that cannot be produced directly. This should

lead eventually to superheterodyne receivers capable of translating optical wavelengths into any desired longer wavelength.

It has been known for some years that if one had a strong enough source of infrared radiation of the right frequency, it would be possible to excite vibrations in a particular species of molecule. Any other molecules that might be present would not be affected. Because the excited molecules would react more vigorously than the others, it should be possible to exert a highly selective control over some chemical reactions. Up to now all available light sources have been far too weak for such possibilities to be taken seriously, but optical masers may eventually make such control a reality.

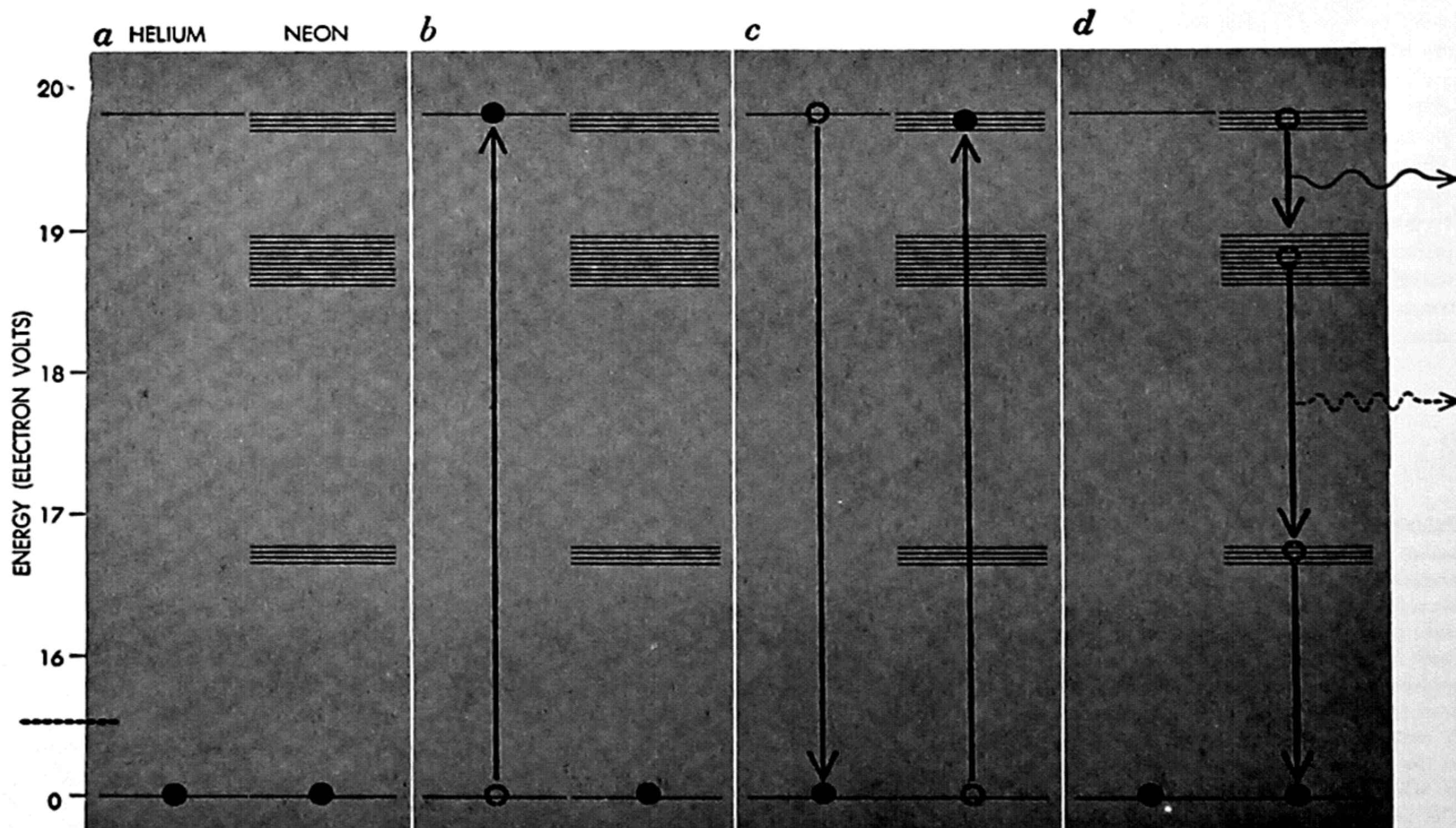
It should be realized that we are talking about a whole family of devices embracing a wide range of frequencies, power ratings and band-widths. The family will include not only oscillators but also amplifiers. One type will be useful for amplifying a light signal that has been weakened by traveling a long distance, perhaps through pipes or through interplanetary space. Another type of amplifier will be able to intensify an entire image—for example, the faint image of a star—that is fed into it.

The list of potential applications of the optical maser could be extended almost indefinitely. With the advent of the optical maser, man's control of light has reached an entirely new level. Indeed, one of the most exciting prospects for workers in the field is that this new order of control will open up uses for light that are as yet undreamed.



CHROMIUM ATOMS (*black dots*) in a ruby maser crystal are "pumped" to higher energy levels and then stimulated to emit photons, producing a maser beam. Atoms in the ground state (*a*) absorb photons (*wavy colored arrows*), which pump them to one

of two energy "bands" (*b*). The atoms give up some of their energy to the crystal lattice and fall to a metastable energy level (*c*). When stimulated by photons from other chromium atoms, they emit photons of a characteristic wavelength and fall to ground state (*d*).



HELIUM AND NEON ATOMS (*black dots*) constitute the active medium of the gas maser. At the start both are in the ground state (*a*). Electron bombardment pumps helium to a higher energy level (*b*). When helium and neon atoms collide, the helium loses its energy to the neon, which is raised to one of four distinct energy

levels (*c*). When stimulated by an outside photon, the neon contributes a photon (*wavy colored arrow at top in "d"*) to the maser beam and falls to one of 10 energy levels. The neon then reverts to the ground state in steps; the photon emitted in the first step (*wavy broken arrow*) does not contribute to the maser beam.

The Author

ARTHUR L. SCHAWLOW is a staff member in the Physical Research Department of Bell Telephone Laboratories. Schawlow was born in Mount Vernon, N.Y., and raised in Canada. His studies at the University of Toronto were interrupted by World War II, during part of which he was engaged in research on microwave guides and antennas. At the end of the war he resumed his graduate work under the direction of M. F. Crawford at Toronto and received his Ph.D. in 1949. He then went to Columbia University as a postdoctoral fellow and research associate. It was there that he began his association with Charles H. Townes, which resulted in the formulation of the basic principles of optical masers. Schawlow joined the Bell Telephone Laboratories in 1951.

Bibliography

INFRARED AND OPTICAL MASERS. A. L. Schawlow and C. H. Townes in *The Physical Review*, Vol. 112, No. 6, pages 1940-1949; December 15, 1958.

THE MASER. James P. Gordon in *Scientific American*, Vol. 199, No. 6, pages 42-50; December, 1958.

POPULATION INVERSION AND CONTINUOUS OPTICAL MASER OSCILLATION IN A GAS DISCHARGE CONTAINING A HE-NE MIXTURE. A. Javan, W. R. Bennett, Jr., and D. R. Herriott in *Physical Review Letters*, Vol. 6, No. 3, pages 106-110; February 1, 1961.

POSSIBILITY OF OBTAINING NEGATIVE TEMPERATURE IN ATOMS BY ELECTRON IMPACT. A. Javan in *Quantum Electronics*, edited by Charles H. Townes, pages 564-571. Columbia University Press, 1960.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

THE MUON

by Sheldon Penman

This fundamental particle has 200 times the mass of the electron but otherwise closely resembles it. Why it should be heavier is a question physicists have sought to answer in recent experiments.

The advance of physics can be compared to that of an army. Salients are thrust out where the opposition is lightest; troublesome areas are bypassed until the necessary forces are available. An example of this pattern is seen in the physicist's method of attack on the puzzle of the mu meson, or muon, a fundamental particle weighing some 200 times as much as an electron. First of the unstable particles to be discovered (with the exception of the free neutron), most tractable in the laboratory because of its relatively long life (2.2 millionths of a second) and the ease with which it can be produced, the muon has been studied in more detail than any other unstable particle. These investigations have shown that it does not, as do other mesons, interact with other particles by "strong coupling," the kind of force that holds together the particles in the nucleus of the atom. Instead the muon appears in every respect except its instability to be a heavy relative of the electron; like the electron, it interacts with the rest of the universe only through the agencies of the electromagnetic force and the so-called weak-interaction force. (The weak interactions give rise to phenomena such as radioactive decay in which an electron is emitted, and the decay of the muon itself.) This raises a baffling question: If the muon is identical to the electron in its interactions, why should it be 200 times heavier? Physicists believe that the mass of a particle is a consequence of its interactions; when two particles display identical interactions, there is no mechanism that can be invoked to explain a difference in their masses.

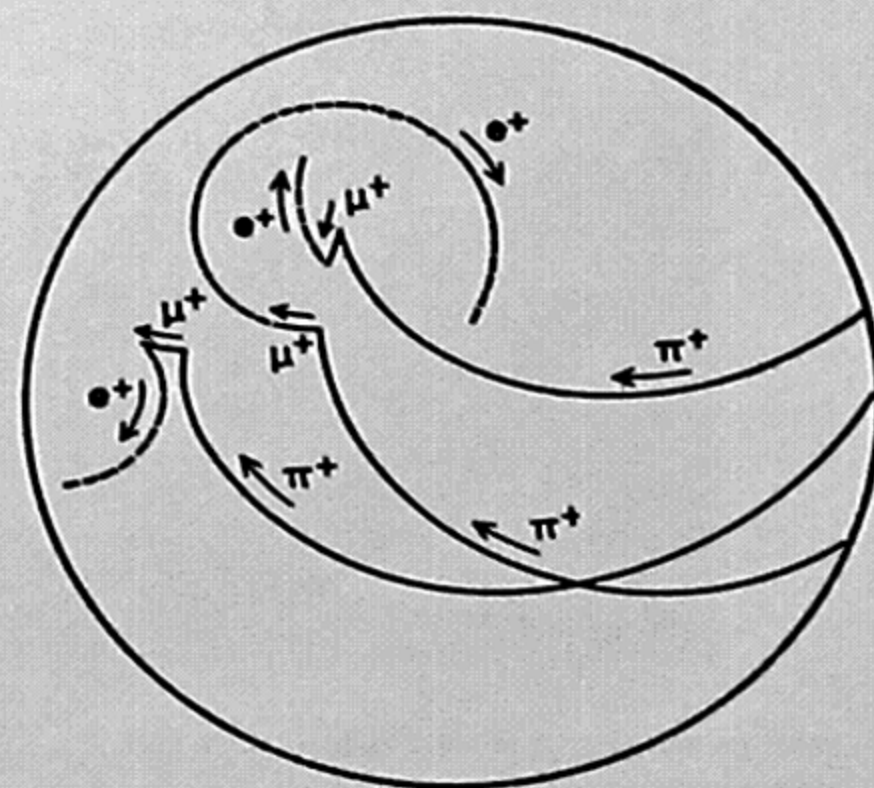
The muon was discovered in 1936 in cloud-chamber photographs of cosmic radiation; the discovery was made by Carl D. Anderson and Seth H. Nedder-

meyer of the California Institute of Technology and independently by J. C. Street of Harvard University. Physicists had been looking for a particle of about this mass since 1935, when its existence was postulated by the Japanese theorist Hideki Yukawa. His calculations had shown that such a particle could explain the enormous strength of the forces that hold together the protons in the nucleus in spite of the mutual repulsion of their positive electric charge. Yukawa reasoned by analogy. The particle that accounts for the electromagnetic force field is the photon, which has no mass. By analogy the nuclear force field should also have its particle, but the particle should have a certain amount of mass. This followed because nuclear forces, unlike electromagnetic forces, extend only a short distance from the nucleus. In fact, the finite range of the nuclear force field indicated that its particle would have a mass about 200 times that of the electron. It soon turned out that particles of this mass—now called mu mesons—existed in embarrassing profusion. The mu mesons in cosmic radiation traveled easily through the atmosphere, penetrated lead plates and could even be detected in deep mines. Such behavior was unbecoming to the hypothetical Yukawa particle. Since it was supposed to give rise to strong forces in the nucleus, it should have interacted readily with the nuclei of any substance through which it passed. It should have penetrated the atmosphere only with difficulty and a lead plate not at all.

Once the low interactivity of the muon had been firmly established, the search for the real Yukawa particle continued. This particle—the pi meson, or pion—was finally found in 1948 as a component in cosmic radiation high in the atmosphere. Its lifetime is so short and it

interacts so readily with nuclei that it rarely penetrates the atmosphere and reaches the ground. Indeed, it meets all the specifications of a particle of the nuclear force field. In addition, it is the parent of the muon, decaying with a mean life of 25×10^{-9} second (25 billionths of a second) into a muon and a neutrino. Were it not for this fortuitous lineage the production of a muon would be a rare event indeed.

The protons of the primary cosmic radiation collide and interact with nuclei in the upper atmosphere. These interactions are "strong"; that is, they are identical with those which produce the nuclear force field. They give rise to various nuclear fragments and a profusion of pions; the pions then decay into muons. A particle that does not enter into strong interactions, such as the muon, could not be manufactured by the direct impact of protons on nuclei. Because particle accelerators produce collisions of this sort they create pions in



LIFE OF THREE MUONS is recorded in three short tracks produced by these parti-

abundance, and these, upon decay, yield an abundance of muons.

In the decade following the discovery of the pion attention was largely diverted from the muon, first by the pion itself and then by the discovery of a bewildering profusion of other particles: K mesons and the particles lambda, sigma and xi. These particles all share a new quantum property called "strangeness," which sets them apart from all the particles discovered earlier. Strangeness, however, is not much stranger than one of the older quantum properties such as "spin"; like matter and energy, it is a property that is

conserved when strange particles enter into strong interactions. When strange particles enter into weak interactions, strangeness is not conserved; no one knows why.

Physicists have found it helpful to show that the particles that enter into weak interactions (putting aside the strange particles) are related by a triangle [see illustration at bottom of next two pages]. The two nucleons (the neutron and the proton) occupy one corner; the electron occupies the second corner; the muon, the third. Experimental evidence indi-

cates—though not yet conclusively—that the particles at the corners of the diagram interact with each other with the same strength.

In the first leg of the triangle the connection between a nucleon and an electron illustrates beta decay, in which, for example, a neutron emits an electron and an antineutrino and is transmuted into a proton. This is also the mechanism by which heavy nuclei decay when they emit an electron, and the rate of decay should be an indication of the strength of the interaction. Many other factors enter into the reaction, however, and

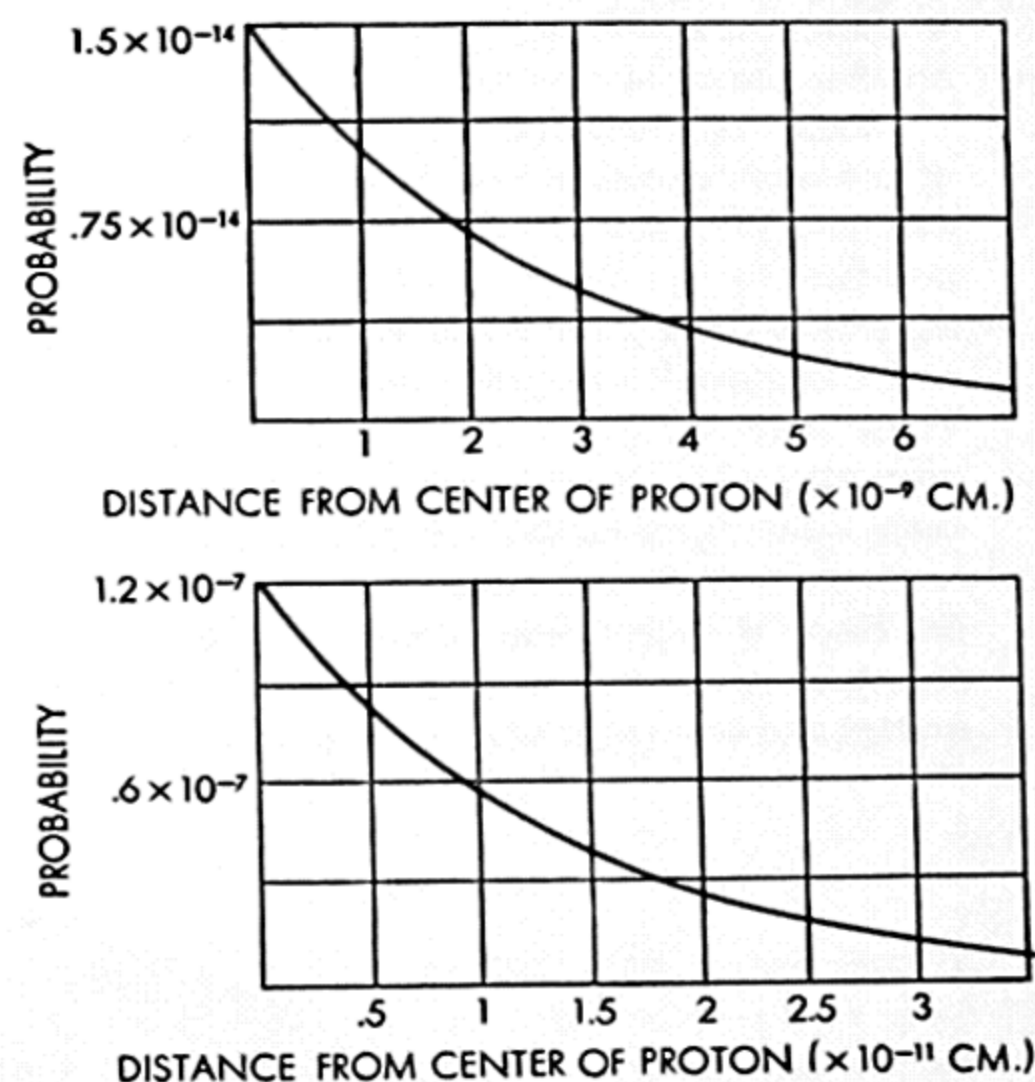


cles in the liquid hydrogen of a bubble chamber. Each of three pi mesons (π^+ in drawing at left) decays into a muon (μ^+) and a

neutrino, which is neutral and leaves no track. Each muon then decays into a positive electron (e^+), a neutrino and an antineutrino.

	ELECTRON	MUON
DATE DISCOVERED	1897	1936
MASS	1	207
MEAN LIFETIME (SECONDS)	INFINITE	2.22×10^{-6}
ELECTRIC CHARGE	-1 (+1)	-1 (+1)
LEPTON NUMBER	1 (-1)	1 (-1)
SPIN	$\frac{1}{2}$	$\frac{1}{2}$
MODES OF INTERACTION	ELECTROMAGNETIC AND "WEAK"	ELECTROMAGNETIC AND "WEAK"
PREDICTED ANOMALOUS MAGNETIC MOMENT	.0011596	.001165
MEASURED ANOMALOUS MAGNETIC MOMENT	$.0011609 \pm .0000024$	$.001145 \pm .000022$

ELECTRON AND MUON differ only in mass, mean lifetime and anomalous magnetic moment (*table at left*). The difference in mass also involves a difference in the probability that either particle will penetrate a nucleus about which it may orbit, as can be seen



by comparing the two "probability curves" at right. The probability of finding an electron at a given distance from the center of a proton in the hydrogen atom is plotted in curve at top; corresponding probability for "mesic" atom (muon and proton), at bottom.

radioactive nuclei actually decay at various rates; some last for much less than a second and others have lasted for the life of the universe. It is a triumph of theory that in the face of rates varying from the positively lethargic to the extremely rapid it is possible to extract information about the fundamental process and show that its strength is always the same.

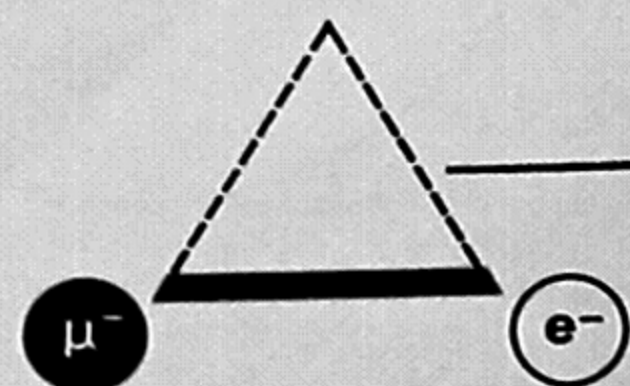
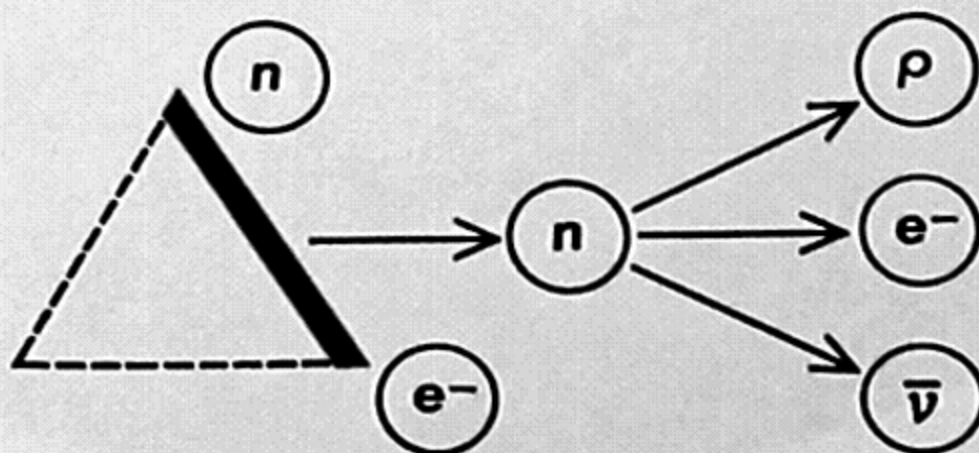
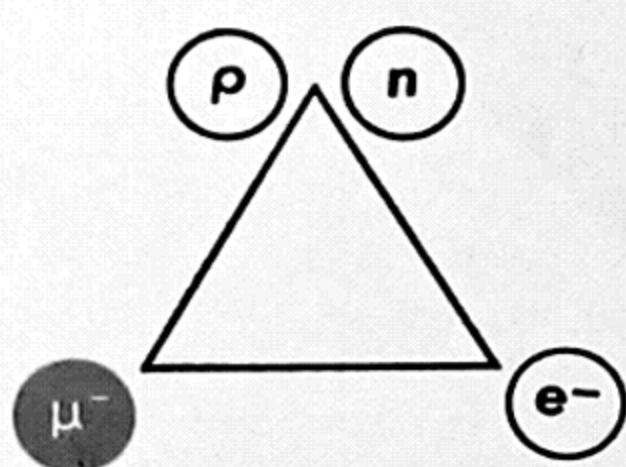
The connection between the muon and the electron in the second leg of the triangle is illustrated by the decay of the muon into an electron, a process that resembles beta decay. Here, however, the outgoing electron is accompanied by not one neutrino but two. When the strength of the interaction of the muon and the electron is compared with that of a nucleon and an electron (on the basis of the rate of decay), there is good agreement.

The third leg of the triangle is the

interaction of muon and nucleon. This interaction is illustrated by the phenomenon of muon capture. When a negative muon comes to rest in any substance, it penetrates the cloud of electrons surrounding a nucleus to which it is attracted by virtue of its negative charge. It then behaves exactly like a heavy electron in that it has only certain energy levels, or orbits, around the nucleus; it cascades down through these permissible levels, emitting an X ray with each transition. This is the way ordinary electrons behave in atoms, and the system of a nucleus and a mu meson is called a mesic atom. When the muon has cascaded as far as it can go, it lives out its remaining life in the state of lowest energy, which is an orbit that is quite close to the nucleus. Because the muon is some 200 times heavier than the electron, its smallest orbit is some 200

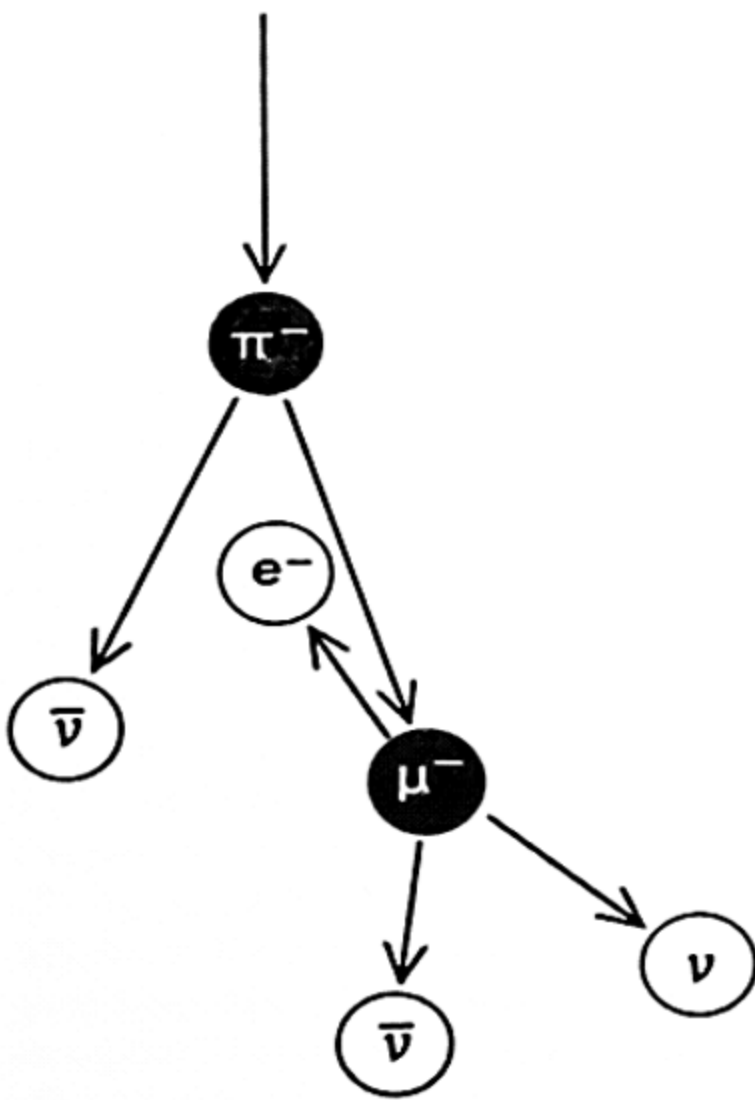
times smaller than that of the electron around the same nucleus.

Modern physics has of course taught us that we should not think of a particle circling the nucleus as a planet circling the sun but rather as a smear of probability, with the particle much of the time actually penetrating the nucleus [see illustration above]. While it is in "orbit" the muon can either decay or interact with a proton in the nucleus, causing the proton to change into a neutron with the emission of a neutrino. As in radioactive decay, the length of time it takes for a muon to interact with a nucleus depends on the particular nucleus involved. Still, also as in radioactive decay, the fundamental strength of the interaction can be deduced. Its value is obtained by studying the rate of capture of muons by many different nuclei. Although the value is

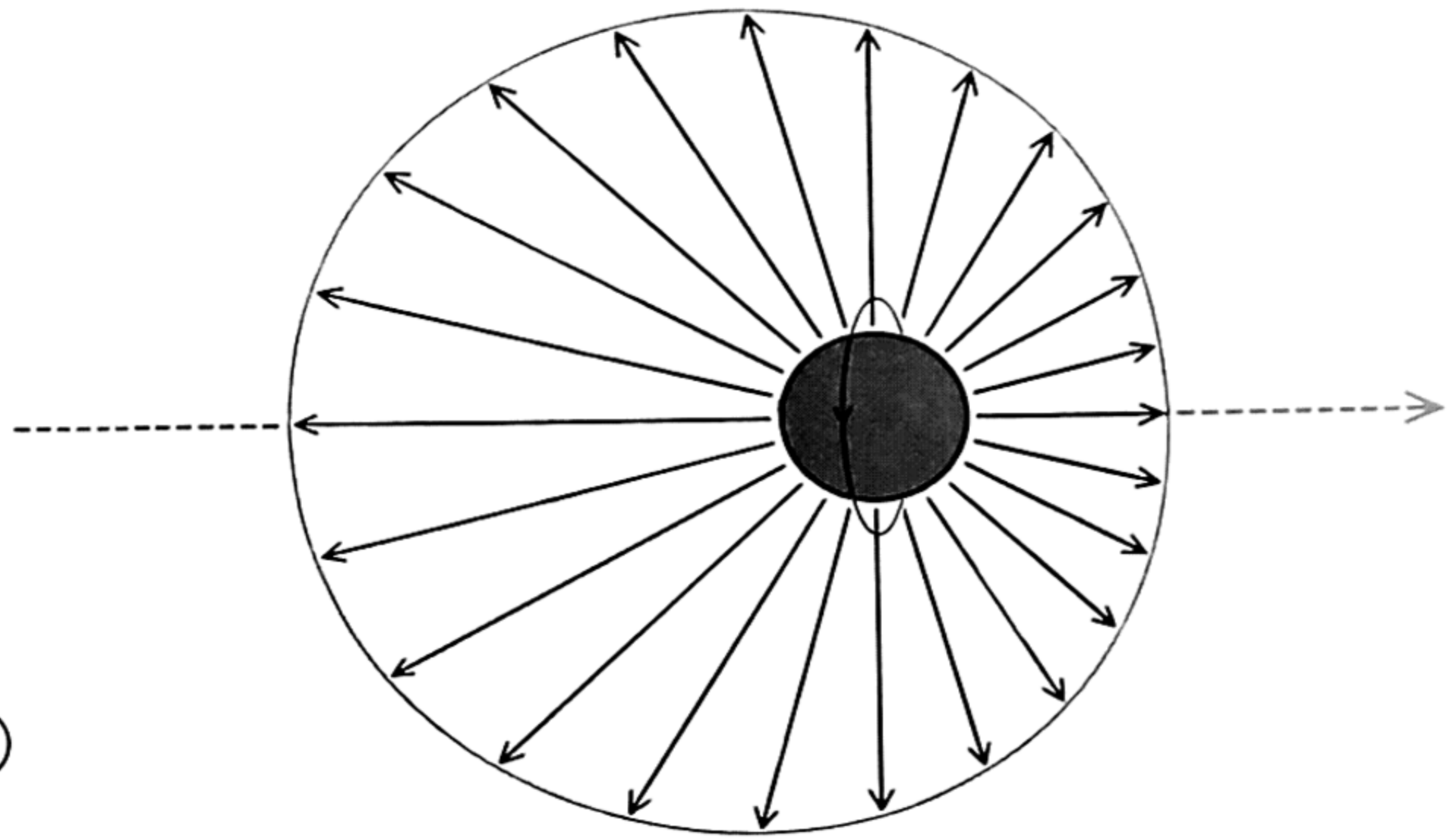


INTERACTION TRIANGLE (*left*) relates particles that enter into weak interactions (leaving aside "strange" particles). In the

first leg (*second from left*) a neutron (n) emits an electron and an antineutrino, becoming a proton (p). Second leg (*third from*



ORIGIN AND DECAY OF MUON (*left*) involves the decay of a pi meson (π^-) and the production of an electron (e^-). The pi meson decays into a muon (μ^-) and an antineutrino ($\bar{\nu}$); the muon decays into an electron, an antineutrino and a neutrino (ν). The muon



(*large colored circle at right*) emits an electron preferentially in a direction opposite that in which its spin axis (*broken colored arrow*) points. The probability of an electron's being emitted in a given direction (*black arrows*) is proportional to arrow lengths.

by no means so well known as the value for radioactive decay, it appears to agree well with that for the "universal" weak interaction. There are two ambitious experiments under way, one at Columbia University and the other at the University of Chicago, to measure the rate at which the nuclei of hydrogen—protons—capture muons. In the capture of muons by protons there is no nuclear structure to contend with and the elemental interaction rate can be found. Although the hydrogen is liquefied and its nuclei are close together, the probability of capture is small and the experimental difficulties are formidable. The results of these experiments should go far toward putting this last leg of the triangle on the same footing as the other two.

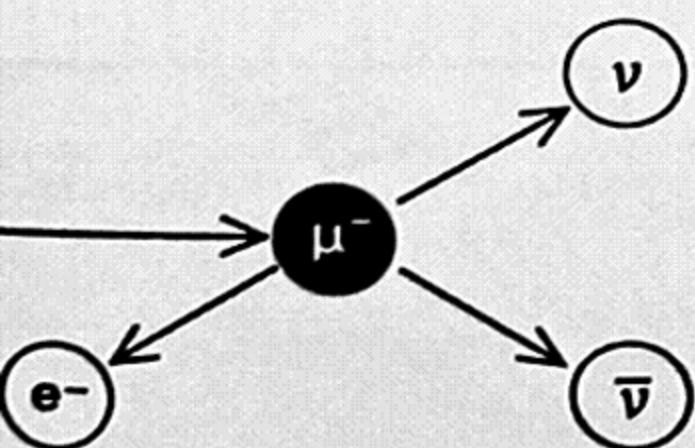
One can therefore conclude that the muon, in so far as the weak interactions go, behaves like a heavy electron. There

are many other similarities. There are, for example, both positive and negative muons, as there are positive and negative electrons. There are no neutral muons and no neutral electrons. The muon's intrinsic angular momentum—"spin"—appears to be $1/2$, as is the spin of the electron. Curiously, the experimental evidence for this fundamental property is not strong, but since so many of the detailed properties of the muon satisfy the theory for a particle of spin of $1/2$, there seems little reason to doubt that $1/2$ it is.

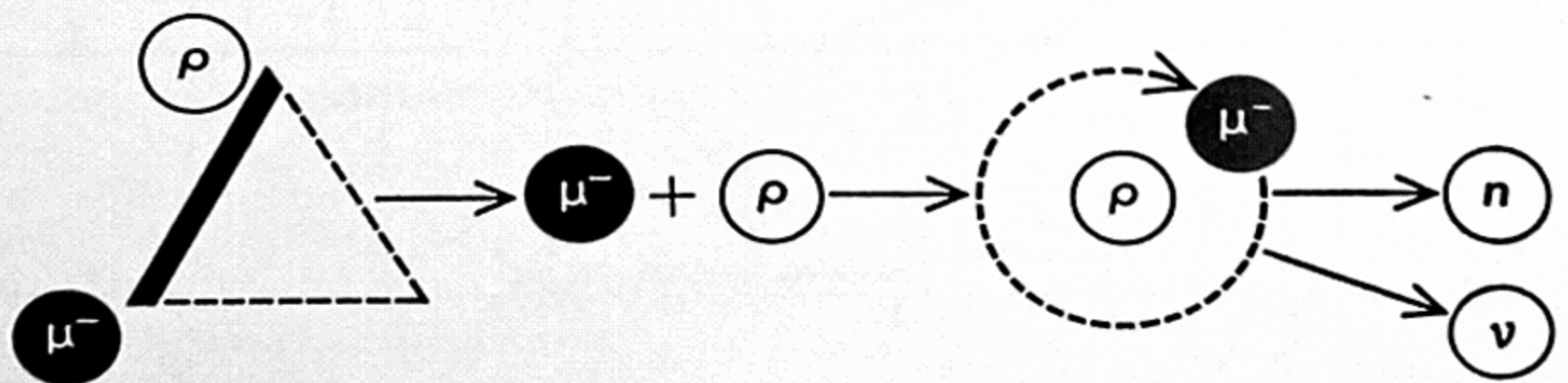
Another similarity between muons and electrons is that they are both "leptons," or light particles. This means that they, together with neutrinos, obey a conservation law stating that the number of leptons in the universe is a constant. This may seem strange inasmuch as there are many processes in which these

particles are created and destroyed. If we count correctly, however, assigning the value of $+1$ to the electron, the negative muon and the neutrino, and a value of -1 to the positron, the positive muon and the antineutrino, then the number of leptons at the beginning and the end of any reaction is the same. During the capture of a negative muon by a nucleus, for instance, a neutrino is emitted. Since both of these particles have a lepton number of $+1$, the lepton equation balances. When a positive muon decays into a positron, a neutrino and an antineutrino, the total number of leptons is -1 both before and after the decay.

This was essentially the state of knowledge in 1957, and it was hoped that more detailed experiments would show properties of the muon that might indicate the origin of its mass. The big im-



left) represents the decay of a muon into an electron, a neutrino and an antineutrino. Third leg (*right*) represents the weak inter-



action in which a proton captures a muon to form a mesic atom. The proton then turns into a neutron and emits a neutrino.

petus to muon research came in that year with the discovery of the nonconservation of parity. The "law" of parity had stated that nature does not distinguish between left and right—in other words, that one cannot perceive a difference between an event in nature and its mirror image. The nonconservation of parity suggested itself to T. D. Lee and C. N. Yang when they were pondering what seemed to be a paradox in the decay schemes of the K meson. (K-meson decay seemed to violate parity and, as we now realize, it did.)

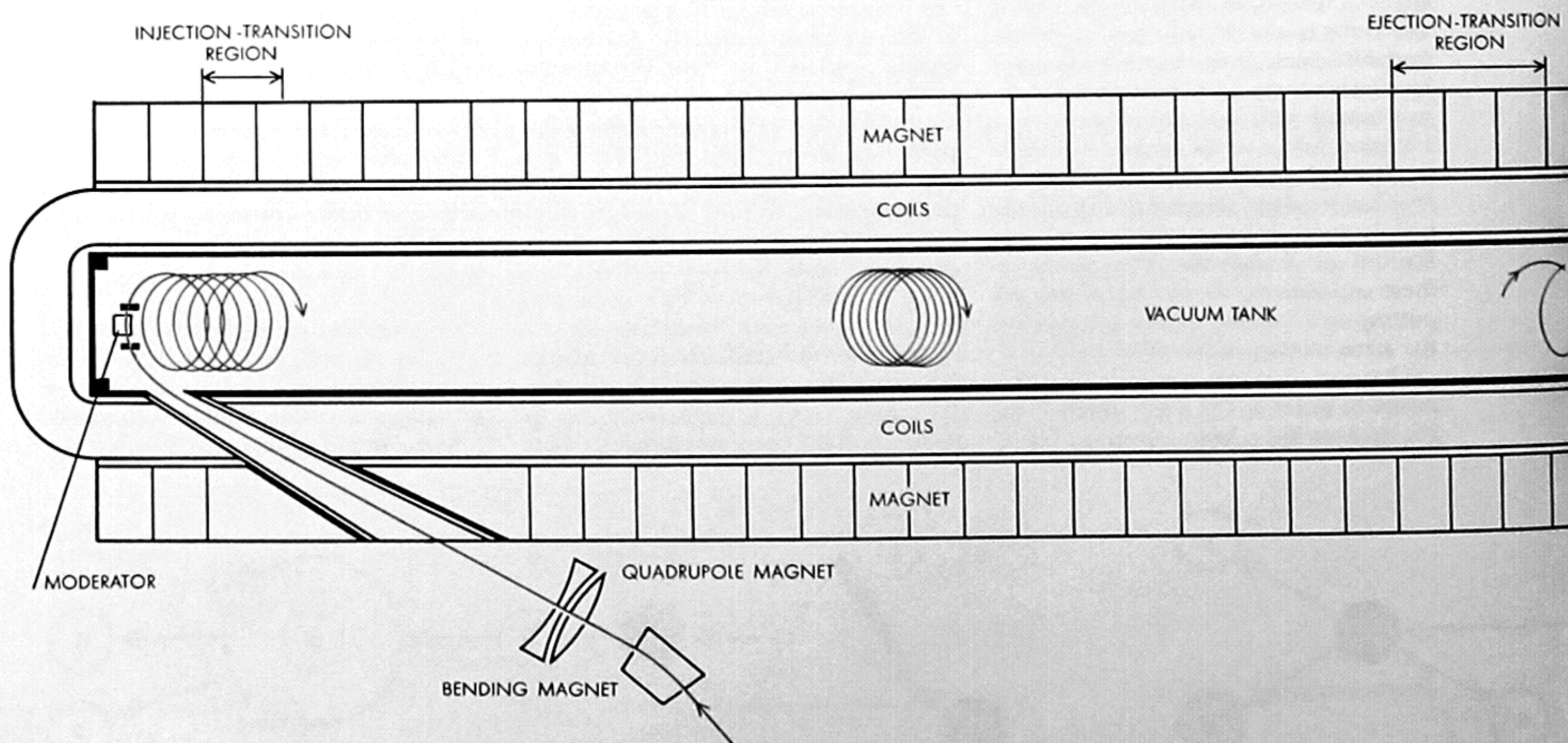
The fall of parity had two important consequences in muon research. First, the muon, because it has spin, can be thought of as having an orientation in space, that is, a direction in which the axis of spin points. It must be realized that to speak of the particle as actually turning is meaningless; this would imply that we could somehow measure the motion of points on its surface, which we cannot. The picture is heuristic—useful because it serves to clarify the nature of the particle. If parity were conserved in the production of muons, the beam of particles emerging from an accelerator would have to contain muons with every possible orientation. Actually it was found that the beams contain particles that tend to point along the direction of travel. Moreover, if parity were

conserved when a muon decayed, there would be no preference in the direction in which its daughter electron was emitted. It happens that, since nature has no interest in conserving parity, the electron is emitted preferentially in a direction opposite to that in which the spin axis of the muon points. Both of these conservation failures were established in the elegant experiment performed at Columbia University by Richard L. Garwin, Leon M. Lederman and Marcel Weinrich.

As a result of these discoveries workers in muon physics now have a powerful means of studying the particle. Not only do the available muons tend to point in one direction but also it is possible to determine that direction by looking for the electron produced in the decay process. To use an analogy, we are no longer trying to handle screws in the dark with heavy gloves; we are being handed the screws neatly aligned on a tray, with a little searchlight on each that indicates the direction of its head. Consequently it is now possible to make measurements of the muon with a precision approaching the precision achieved in the investigation of atoms.

Almost immediately after the discovery of the breakdown of parity, a series of experiments was initiated at Columbia University, the University of Chicago

and most recently at C.E.R.N. (the European Organization for Nuclear Research) to study the magnetism of the muon. Every charged particle possessing spin behaves as though there were a bar magnet lying along its spin axis; the strength of this bar magnet is called its magnetic moment. In the case of the electron, electromagnetic theory can predict the value of the magnetic moment with high precision. It is of great interest to measure the magnetic moment of the muon to determine if the theory that works so well for the electron is applicable. A deviation from the predicted value could give an important clue to the structure of the muon and hence to the origin of its mass. In the proton, for example, the value of the magnetic moment is quite different from the value that would be predicted purely on the grounds of its electromagnetic interactions. In modern particle physics the proton is regarded as continuously emitting and absorbing mesons; because the proton interacts by strong coupling the mesons are pions. These pions are not physically observable; they are called "virtual." Nevertheless, they produce circulating electric currents that alter the magnetic moment of the proton. Since the muon does not enter into strong interactions, any departure from its predicted magnetic moment could



C.E.R.N. APPARATUS, used in so-called $g-2$ experiment, employed a vacuum chamber 20 feet long between the poles of an 85-ton magnet. Theory says that when a muon is forced to orbit in a magnetic field, its spin axis should precess, or turn at a slightly

faster rate than its orbiting rate. The amount of precession is indicated by flight direction of electrons emitted when muons decay in target (far right; see also illustrations on next two pages). After muons (colored line) enter the vacuum chamber they are slowed

perhaps be attributed to some heretofore unknown feature of the muon's structure or even to an interaction with the field of a particle still undiscovered. As an alternative explanation, a departure in the magnetic moment of the muon could indicate a breakdown in the laws of electromagnetism at very short distances. This could show up in the muon and not in the electron because the muon's greater mass implies that its interactions with the electromagnetic field take place over smaller regions of space. Theorists have speculated for a long time that space, like energy, may not be indefinitely divisible and that a shortest length may yet be found that will limit the applicability of the electromagnetic laws as we know them. While there is no evidence for such a fundamental quantum of length, the magnetic moment of the muon gives the physicist a new tool for probing the question.

The unit of strength of the magnetic moment of a particle is the magneton. Since the definition of the magneton involves mass, the unit is different for each particle. The value of the moment is often expressed in terms of a "g-factor," which dates back to the hoary lore of atomic physics. The g-factor is equal to the ratio of the actual magnetic moment in magnetons to a value that is half a magneton. A major triumph of

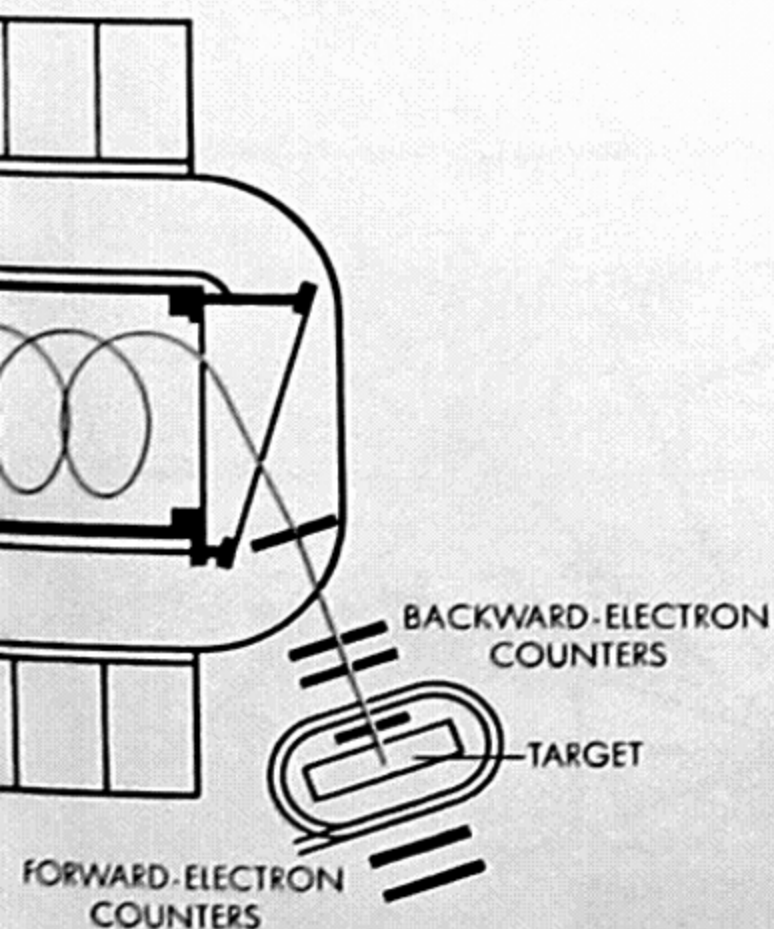
P. A. M. Dirac's celebrated formulation of quantum mechanics was its prediction that the g-factor of the electron was equal to 2, which agreed with the experiments of that time.

Shortly after World War II more precise experiments, performed at Columbia University by Polykarp Kusch, indicated that the g-factor of the electron actually differed from 2 by about one part in 1,000. The departure became known as the anomalous magnetic moment of the electron. Before long Julian Schwinger of Harvard refined the equations of quantum electrodynamics so that the theoretically predicted value of the g-factor once again agreed with experiment. Small though the anomaly is, it is crucial to an understanding of the interaction of charged particles with the electromagnetic field. The anomaly arises because the electron is constantly emitting and reabsorbing virtual photons, which in themselves are unobservable but which give rise to measurable effects.

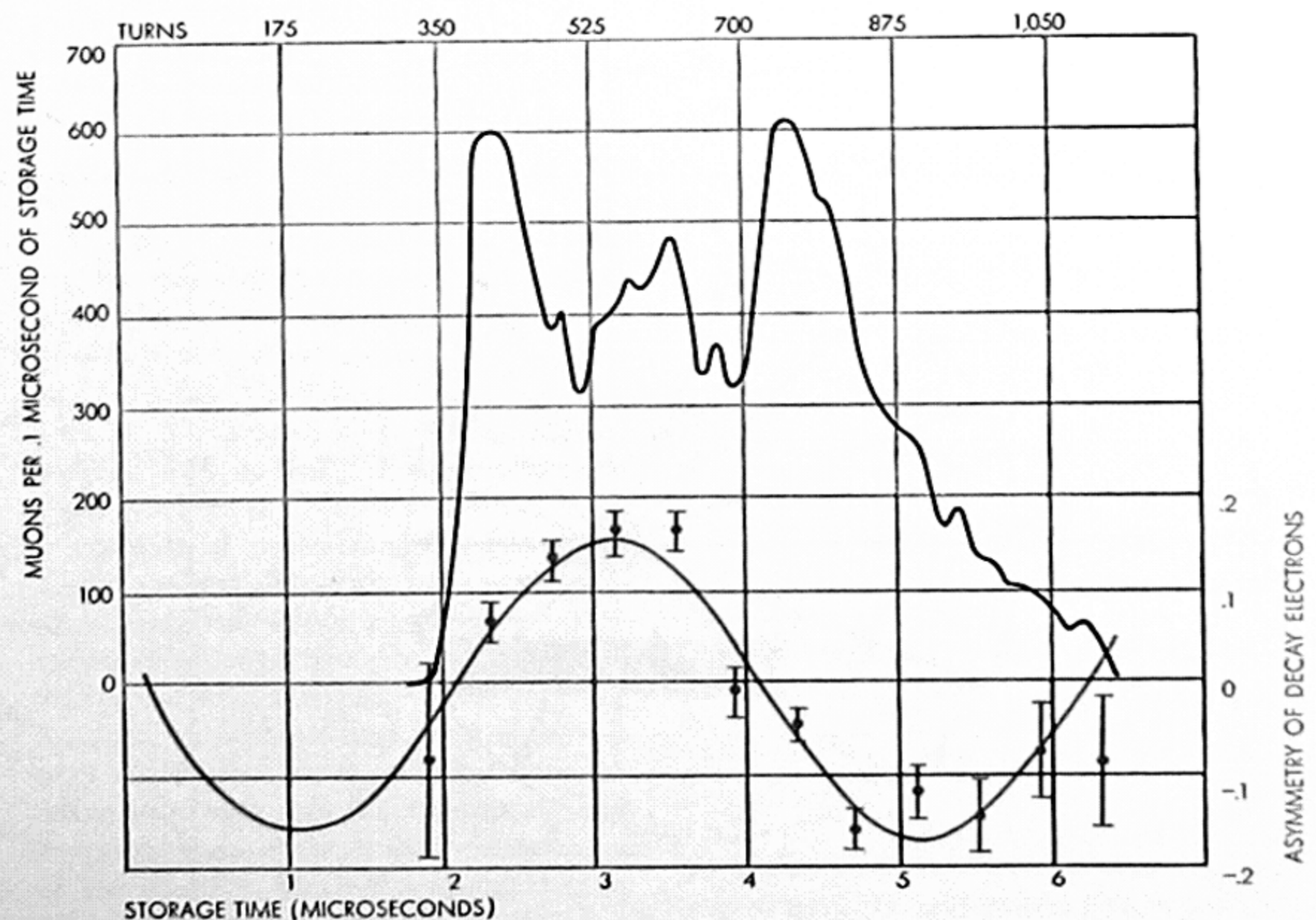
The goal of several experiments undertaken since 1957 has been to measure the g-factor of the muon as precisely as possible. There have been two approaches. The first and simplest is to measure the magnetic moment directly. This can be done with high precision, and experiments are now approaching an accuracy

of one part in 100,000. Theory, however, does not predict the moment directly; it predicts only the g-factor, and this prediction requires knowing the mass of the muon. Although the mass of the muon has been measured far more precisely than that of any other unstable particle, the measurement is still limited in accuracy to about one part in 10,000. As a result the experimental value for the g-factor is limited to the same order of accuracy. Since the deviation of the g-factor from 2 is only one part in 1,000, a measurement accurate only to one part in 10,000 leaves an uncertainty of 10 per cent, which is much too large to provide a meaningful test of the predicted value.

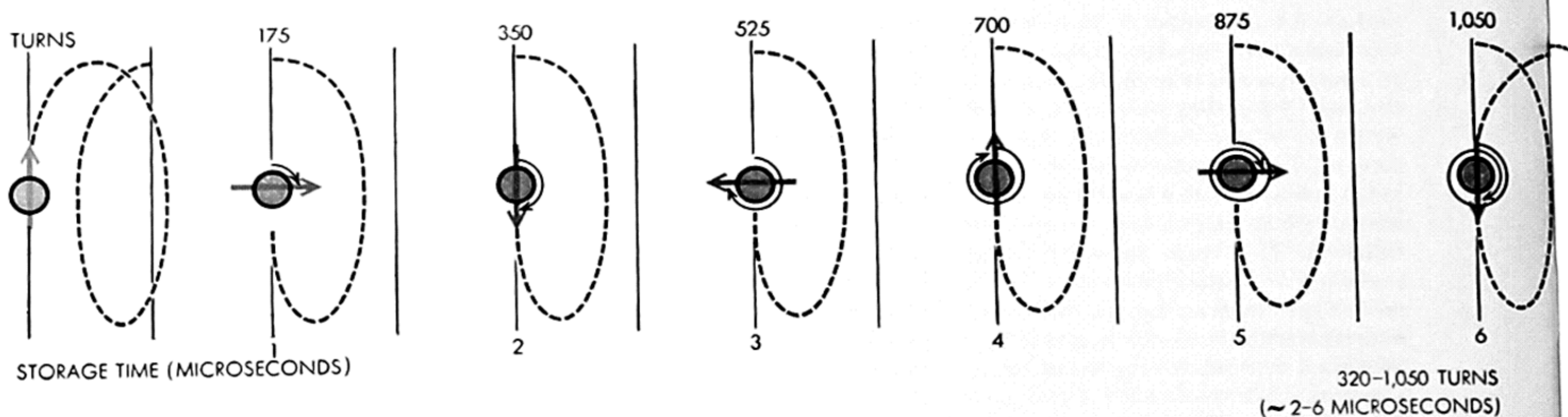
The direct measurement of the muon's magnetic moment has nonetheless been a useful and satisfying achievement. The most recent measurement was carried out at Columbia University by D. P. Hutchinson, J. Menes, A. Patlach, G. Shapiro and the author. To make the measurement one observes the rate at which the spin axis of the muon precesses, or turns, when the particle is brought to rest in a magnetic field. The rate of precession depends on the magnetic moment and the strength of the field. The muons are created by the decay of pions produced in a cyclotron; when the pions decay in flight, they give rise to muons whose spin axes are point-



down by a moderator and begin orbiting. Variations in magnetic field force muons to "walk" from left to right while describing their orbits.

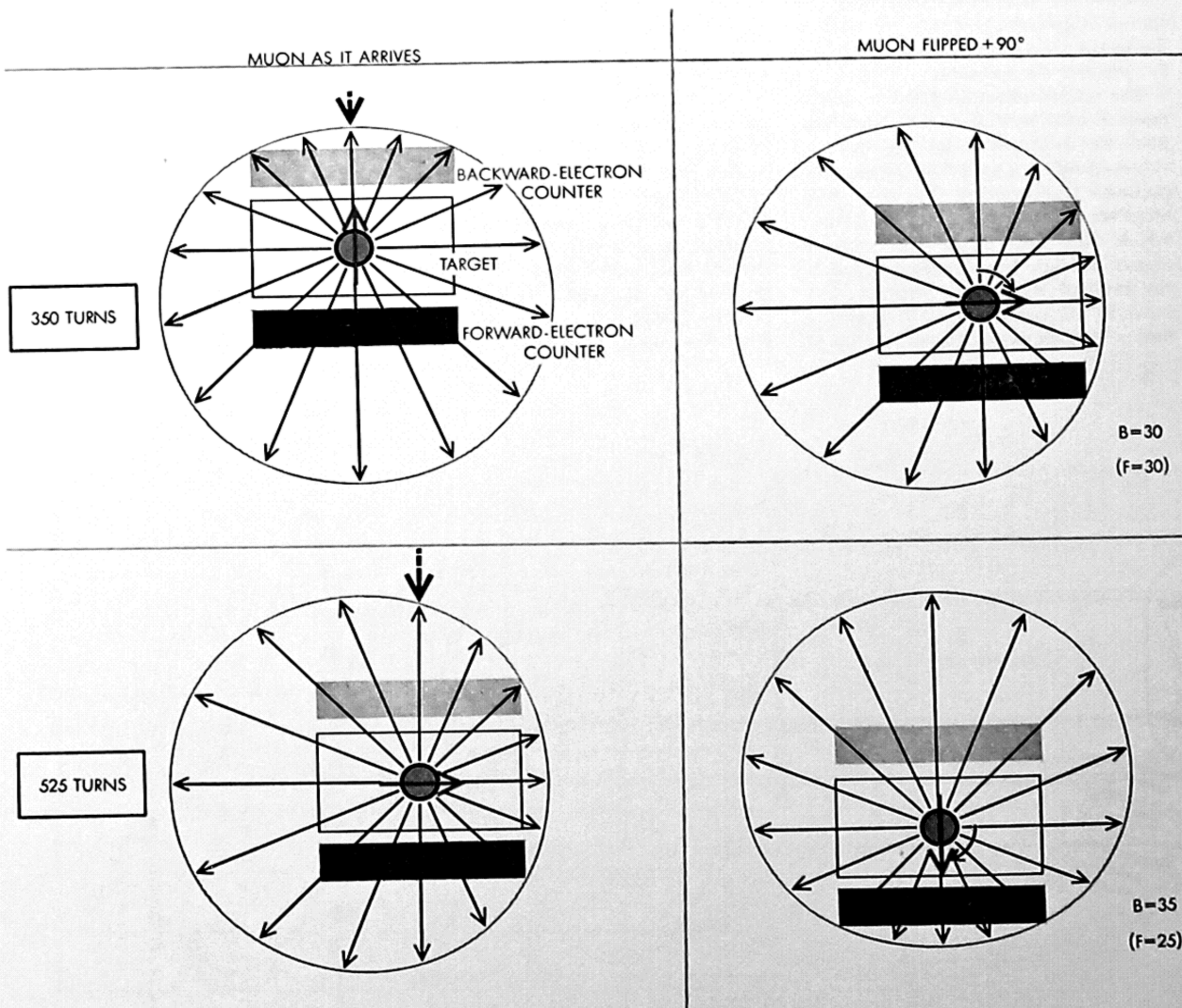


PLOT OF C.E.R.N. RESULTS shows length of time muons spent in the magnet and number of turns they made (upper curve) and the asymmetry of the electrons emitted when muons decayed in the target (vertical colored lines). The sinusoidal curve that best fits the experimental points determines the muon's anomalous magnetic moment: $.001145 \pm .000022$.



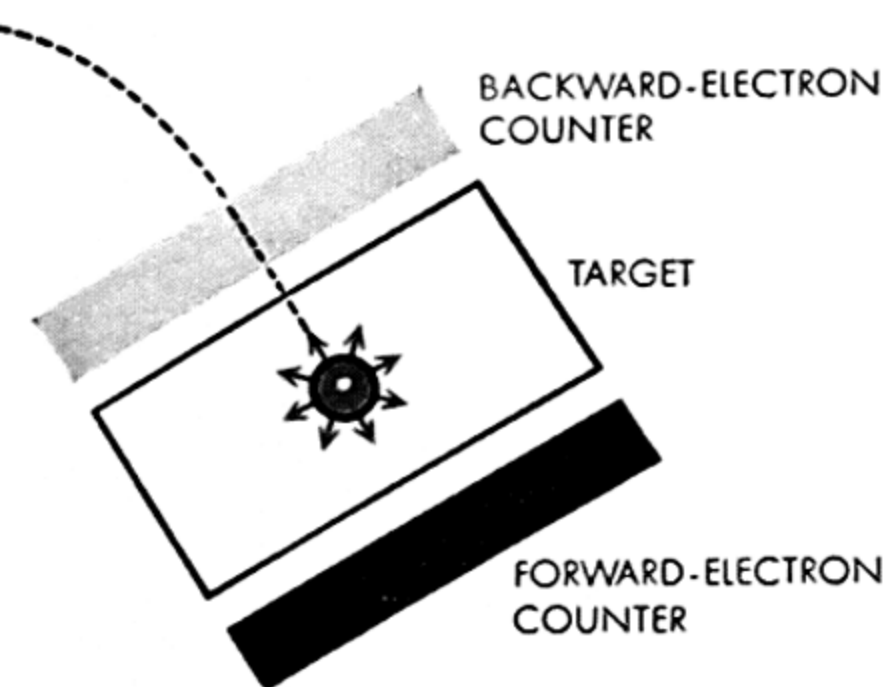
PRECESSION OF MUON SPIN AXIS (colored arrow) in C.E.R.N. apparatus (see illustration on preceding two pages) is the basis for determining the anomalous magnetic moment of the muon. At start

(left) the spin axis points in the direction in which the muon is traveling. The axis has precessed 90 degrees after 175 turns (second from left) and 540 degrees after 1,050 turns (right). Precession is



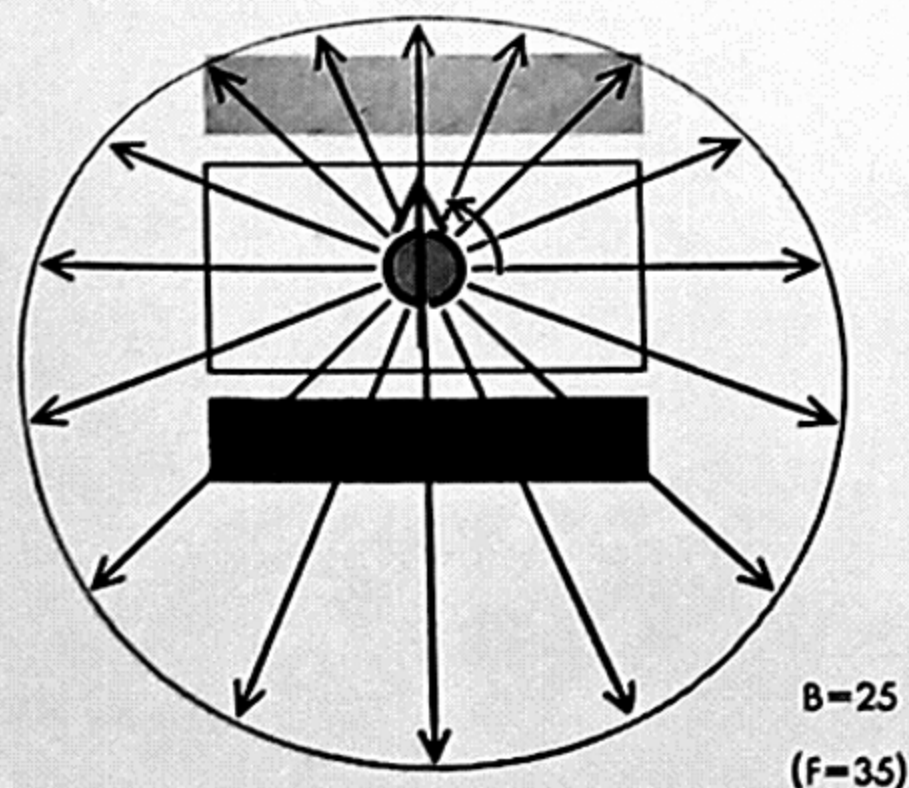
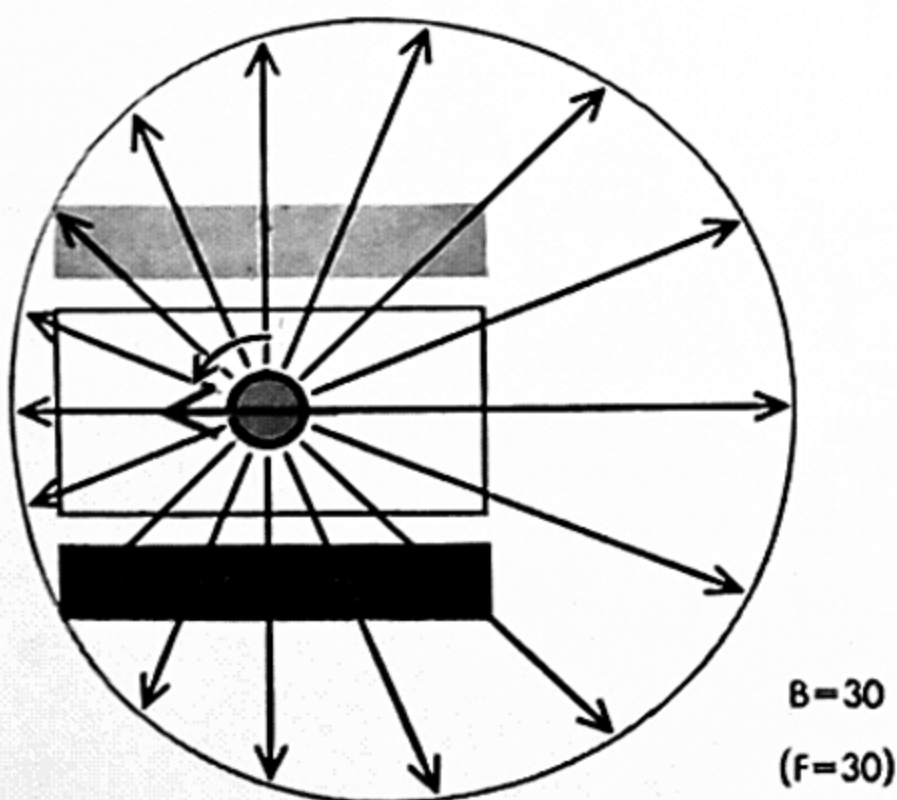
ELECTRONS EMITTED IN MUON DECAY provide the clue to muon precession rate. Muon that has made 350 turns in magnet enters target with spin axis pointing backward. Black arrows radiating from muon are proportional in length to probability of

electron emission. To eliminate systematic errors, electron counts are made on groups of muons magnetically flipped +90 degrees and on others flipped -90 degrees. Counts (backward) should be identical for muons making 350 turns (or 700 or 1,050 turns).



calculated from the "storage" time and the direction in which electrons are emitted when muons decay in the target (see below).

MUON FLIPPED - 90°



For muons making, say, 525 turns flipping will produce different counts (e.g., 35 v. 25 in backward direction). Forward-electron counts (parentheses) show similar behavior.

ing predominantly forward. The muons are then stopped by a target and come to rest with their spin axes perpendicular to the surrounding field of a large electromagnet. Immediately the muons begin their precession.

Electron counters placed around the target record when and in what direction muons decay into electrons. Since the electrons are preferentially emitted in a direction opposite that in which the spin axes of the muons are pointing at the moment of decay, the electrons serve as a searchlight beam that sweeps around at the rate of precession. It is true that the searchlight is turned on only at the instant of decay; nevertheless, if many muons each of which decays after a different length of time are observed, a satisfactory reconstruction of the precession can be made. Since the experiment consists of counting how many turns are made in a given interval of time, and since the muon has a mean life of only 2.2 millionths of a second, it is desirable to use a strong magnet to achieve as high a precession rate as possible. Our experiment used a magnetic field of 15,000 gauss, which gives rise to a precession rate of nearly 200 million times per second. The number of turns can be counted by electronic equipment that compares the muon precession rate with a voltage varying at a precisely known frequency. The equipment is so complicated that systematic errors can easily creep into the measurements unless it is calibrated with great care. Our principal calibration scheme employed an electronically generated artificial signal that mimicked the behavior of muons precessing at a high and accurately known rate.

The most accurate method of measuring the mass of the muon—which must be combined with the magnetic moment to obtain the g-factor—was first devised at Columbia University five years ago by L. J. Rainwater. The precision of the method has recently been refined by other workers at Columbia and Chicago. It consists of measuring the energy of X rays (which, like light, consist of photons) emitted when a negative muon cascades down through its allowable orbits around a nucleus. The energy follows the same laws that govern the energy of the photons emitted by electrons when they make similar transitions, and this energy depends directly on the mass of the particle. Since the muon is more than 200 times heavier than the electron, the photons it emits are correspondingly more energetic. Unfortunately the few hundred photons per second produced

by mesic atoms are far too few to activate the conventional X-ray spectrometers that could measure their energy.

There is, however, a way around this difficulty. It happens that when X rays are directed at an absorbing target, rays having a certain narrow range of energies are absorbed much more strongly than rays of slightly less energy. These sharp changes in absorption, known as edges, occur when an X ray has just enough energy to knock an electron occupying a particular energy level clear out of an atom. By coincidence X rays emitted by a muon in a certain step of its cascade from orbit to orbit around a phosphorus nucleus have an energy corresponding to an edge in lead known as the K edge. The edge is not absolutely sharp but changes over a narrow energy interval, which has been accurately measured. By allowing X rays from phosphorus mesic atoms to be absorbed in lead one can determine just where in this energy interval the X-ray line is located. From this it is possible to calculate that the mass of the muon is $206.76 \pm .02$ times the mass of the electron. While this is a remarkably precise value as mass measurements go, it still leaves a tantalizing uncertainty of 10 per cent in the value of the g-factor.

Quite a different way to get at the g-factor was taken in a remarkable experiment at C.E.R.N. that required the best part of three years to carry out from conception to final completion early this year. The investigators were a five-

ASYMMETRY_(B) =

$$\frac{B_{+90^\circ} - B_{-90^\circ}}{B_{+90^\circ} + B_{-90^\circ}}$$

ASYMMETRY_(B) (350 TURNS) =

$$\frac{30 - 30}{30 + 30} = \frac{0}{60} = 0$$

ASYMMETRY_(B) (525 TURNS) =

$$\frac{35 - 25}{35 + 25} = \frac{10}{60} = .17$$

ASYMMETRY EQUATION uses electron counts from C.E.R.N. experiment to provide values for asymmetry curve on page 637. Examples show equation solved for 350 turns and for 525 turns using sample backward counts (B) found in the illustration at left.

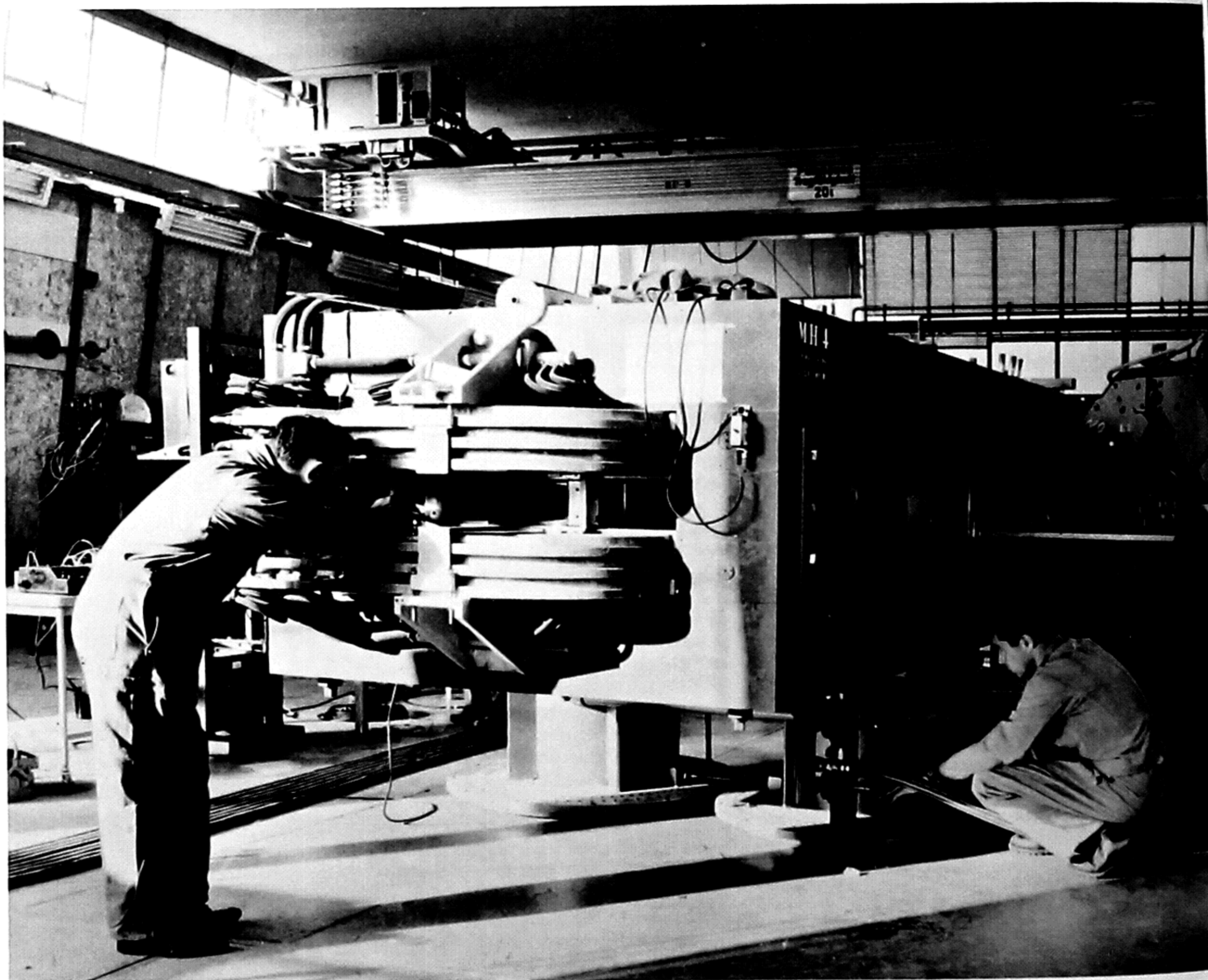
nation team consisting of G. Charpak and T. Muller (France), A. Zichichi (Italy), J. C. Sens (Netherlands), F. J. M. Farley (United Kingdom), Richard L. Garwin and V. L. Telegdi (U. S.). In this experiment the lack of knowledge of the muon mass was circumvented by measuring directly the deviation of the g -factor from a value of 2, which results in its being called the " g minus 2" experiment. The experiment was suggested in part by a $g-2$ experiment recently carried out on the electron at the University of Michigan by A. A. Schupp, R. W. Pidd and H. R. Crane.

When a charged particle moves

through a magnetic field so that its path is perpendicular to the direction of the magnetic field, it is subjected to a force that makes the trajectory of the particle curve. In a uniform field this orbit will be a circle, and the time required for a particle to complete one turn around the orbit will depend only on the strength of the field and the charge and mass of the particle. The time is independent of velocity (at least until the particles attain velocities close to that of light) because the diameter of the particle's orbit increases with velocity so as to keep the time needed for one revolution constant. This is the principle on which the cyclo-

tron operates, and the frequency with which the particle sweeps out its circular orbit is called the cyclotron frequency.

We have already seen that a particle possessing a magnetic moment will precess in a magnetic field. A particle with a g -factor of exactly 2, and with a spin axis pointing in the direction in which the particle is moving when it enters a magnetic field, will precess just enough for its spin to remain pointed along its orbiting path. If the g -factor is different from 2, the cyclotron and precession frequencies will no longer be identical and the spin axis will gradually precess with respect to the trajectory. If



G-2 APPARATUS AT C.E.R.N. was used to determine the muon's anomalous magnetic moment. The man at left is looking between

the magnet coils into the region where muons orbit. The muons are produced when pions, created by a cyclotron, decay in flight.

one stops a particle after it has spent a known amount of time in a known magnetic field and measures how large an angle the spin axis makes with the direction in which the particle is traveling, one then knows the amount by which the g -factor deviates from 2.

The principle of the experiment is easily stated. The actual physical realization was quite another matter. Difficult as the electron $g-2$ experiment was, the muon $g-2$ measurement was even harder; indeed, many physicists believed it would be impossible. Compared with electrons, muons are available in tiny numbers and their lifetime is extremely

short. Moreover, the relatively large mass of the muon calls for strong magnetic fields. The more turns the muon is forced to make, the greater will be its precession from its line of flight and the greater will be the accuracy of the measurement. In the C.E.R.N. experiment a field of 16,000 gauss was used and muons spent from two millionths to six millionths of a second orbiting in the field.

A magnetic field capable of containing particles that enter with the widely varying momentum, direction and position that characterize a muon beam must be carefully designed and painstakingly adjusted. The C.E.R.N. magnet weighs 85 tons and encloses a vacuum chamber nearly 20 feet long [see illustration on pages 636 and 637]. Muons enter one end of the chamber and are slowed by passing through a beryllium block so that they begin curving into the desired orbit. To prevent the muons from simply describing a circle and striking the block again the magnetic field is shaped so as to produce a "walking" orbit. This is done by making the magnetic field stronger along one side of the vacuum chamber than along the other; where the field is stronger the muon is forced to make a tighter turn and therefore is displaced sideways after each loop of 360 degrees. The field is so shaped that the orbits first walk about two centimeters per turn to get them away from the beryllium quickly. This is followed by a region in which the orbits are packed tightly together (to minimize the length of the chamber), and finally the steps are increased again to give the muon a shove to get it past the fringe field of the magnet. In addition the magnetic field must provide vertical focusing so that particles whose motion is not strictly horizontal will not strike the magnet faces. Upon leaving the magnetic field the muons enter a target where they stay until they decay and emit an electron. Many muons, of course, decay en route and never reach the target.

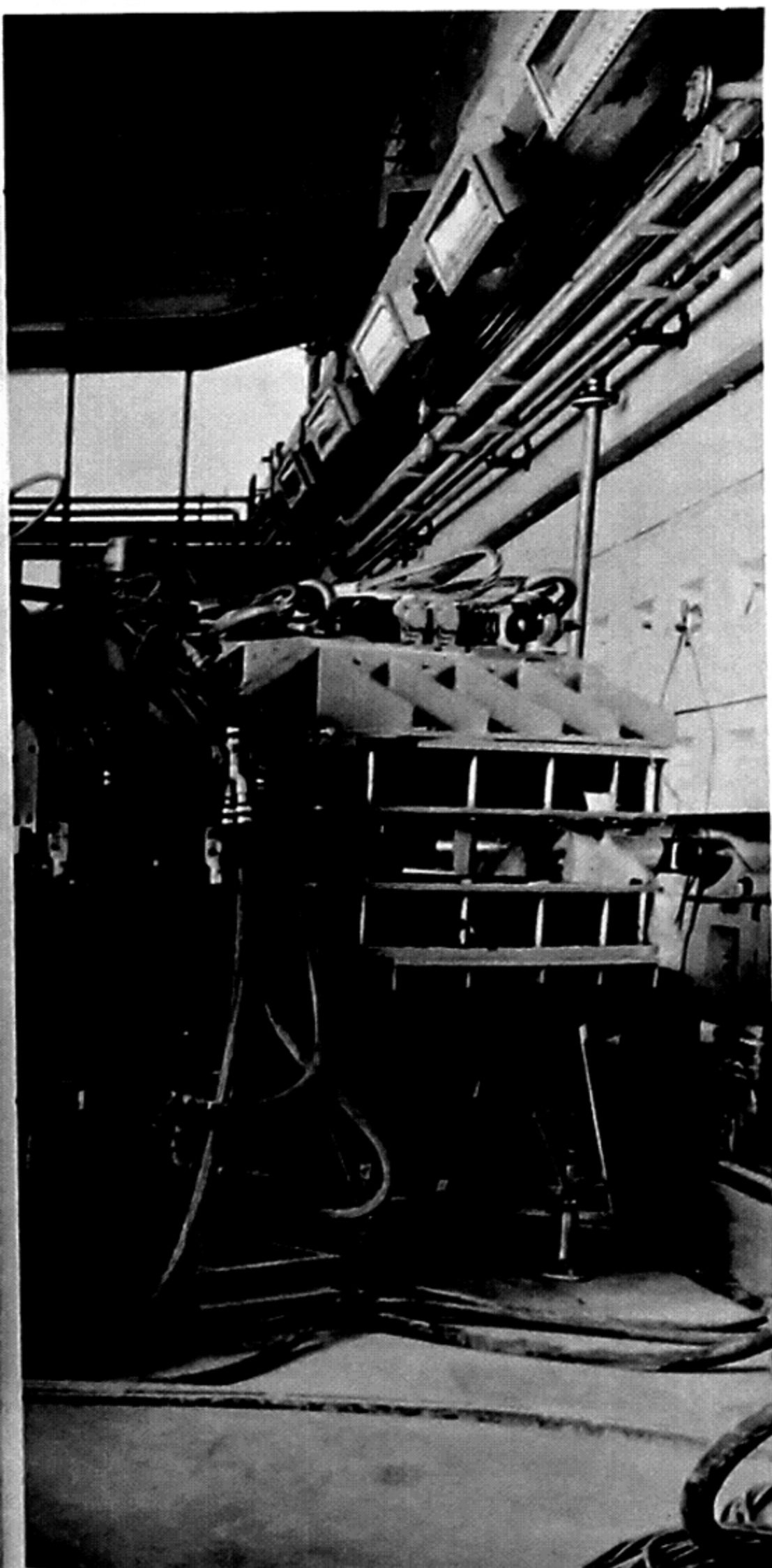
Two measurements must be provided by the experiment: the length of time each muon spends inside the magnetic field and the direction of the electron emitted when the muon decays in the target. The "storage" time in the magnetic field indicates the number of orbits the muon has made. Muons that happen to take large "steps" can walk through by making as few as 320 or so orbits; those taking very short steps may require over 1,000 orbits. Obviously when two muons enter the magnetic field in

rapid succession, there is no way to tell which will move through faster and emerge first. The solution to this is to count only muons spaced far enough apart so that it is physically impossible for the second to overtake the first.

To obtain the second measurement one could try to determine the precise angle at which every decay electron was emitted. Even if this were feasible, however, the electron would indicate only the probable orientation of the muon that had ejected it. Since the direction of emission is probabilistic, it is simpler to count only the decay electrons that emerge in a fixed direction from successive groups of muons. First a little trick is performed. The muons in half the groups, after coming to rest in the target, are turned 90 degrees clockwise by pulsing a small magnetic field around the target. The muons in the other groups are turned 90 degrees counterclockwise. The two directions of pulsing can be expected to yield for each group a slightly different number of electron counts. Turning the muons in this fashion removes the systematic errors that would attend an effort to use two (or more) sets of detectors to count electrons in two (or more) directions [see illustrations on pages 638 and 639].

The experiment shows that the g -factor of the muon is $2.001145 \pm .000022$. The theoretical prediction is 2.001165. Experiment therefore confirms, to an accuracy of 1 per cent in the anomalous part of the g -factor, that the muon behaves exactly like a heavy electron. This result also implies that there is no breakdown in the laws of electromagnetism down to distances of 7×10^{-14} centimeters and no fundamental quantum of length greater than 2×10^{-14} cm. In addition it appears that any hope of explaining the mass of the muon by heretofore undetected interactions with fields other than the electromagnetic (such as fields produced by still undiscovered particles) must be abandoned.

The mystery of the muon mass has deepened and at the moment there are no very helpful suggestions as to where physicists can turn for enlightenment. The best hope seems to lie in scattering experiments using the higher energy muons from the new 30-billion-electron-volt accelerators at C.E.R.N. and at the Brookhaven National Laboratory. There is always a chance that something new will turn up in such ultraenergetic collisions. For the time being, however, the muon itself qualifies as a "riddle wrapped in a mystery inside an enigma."



Before entering the big magnet the muon beam is focused by smaller magnet at right.

The Author

SHELDON PENMAN is assistant professor of physics in the Enrico Fermi Institute for Nuclear Studies at the University of Chicago. He took a B.S. in physics while working in a Naval Ordnance laboratory, did graduate work at Columbia University and acquired his Ph.D. there in 1958. Penman, who joined the Enrico Fermi Institute in 1959, has been engaged in research on muons (mu mesons) for the past three years. At present he is doing experimental work on the capture of muons by atomic nuclei.

Bibliography

ACCURATE DETERMINATION OF THE μ^+ MAGNETIC MOMENT. R. L. Garwin, D. P. Hutchinson, S. Penman and G. Shapiro in *The Physical Review*, Vol. 118, No. 1, pages 271-283; April 1, 1960.

THE EXPERIMENTAL STATUS OF THE WEAK INTERACTIONS OF NON-STRANGE PARTICLES. V. L. Telegdi in *Proceedings of the 1960 Annual International Conference on High Energy Physics at Rochester*, pages 713-725; 1960.

MEASUREMENT OF THE ANOMALOUS MAGNETIC MOMENT OF THE MUON. G. Charpak, F. J. M. Farley, R. L. Garwin, T. Muller, J. C. Sens, V. L. Telegdi and A. Zichichi in *Physical Review Letters*, Vol. 6, No. 3, pages 128-132; February 1, 1961.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

GAS CHROMATOGRAPHY

by Roy A. Keller

A simple analytical method sharply separates components of complex mixtures. When used to characterize perfumes and flavors, its sensitivity rivals that of the human nose.

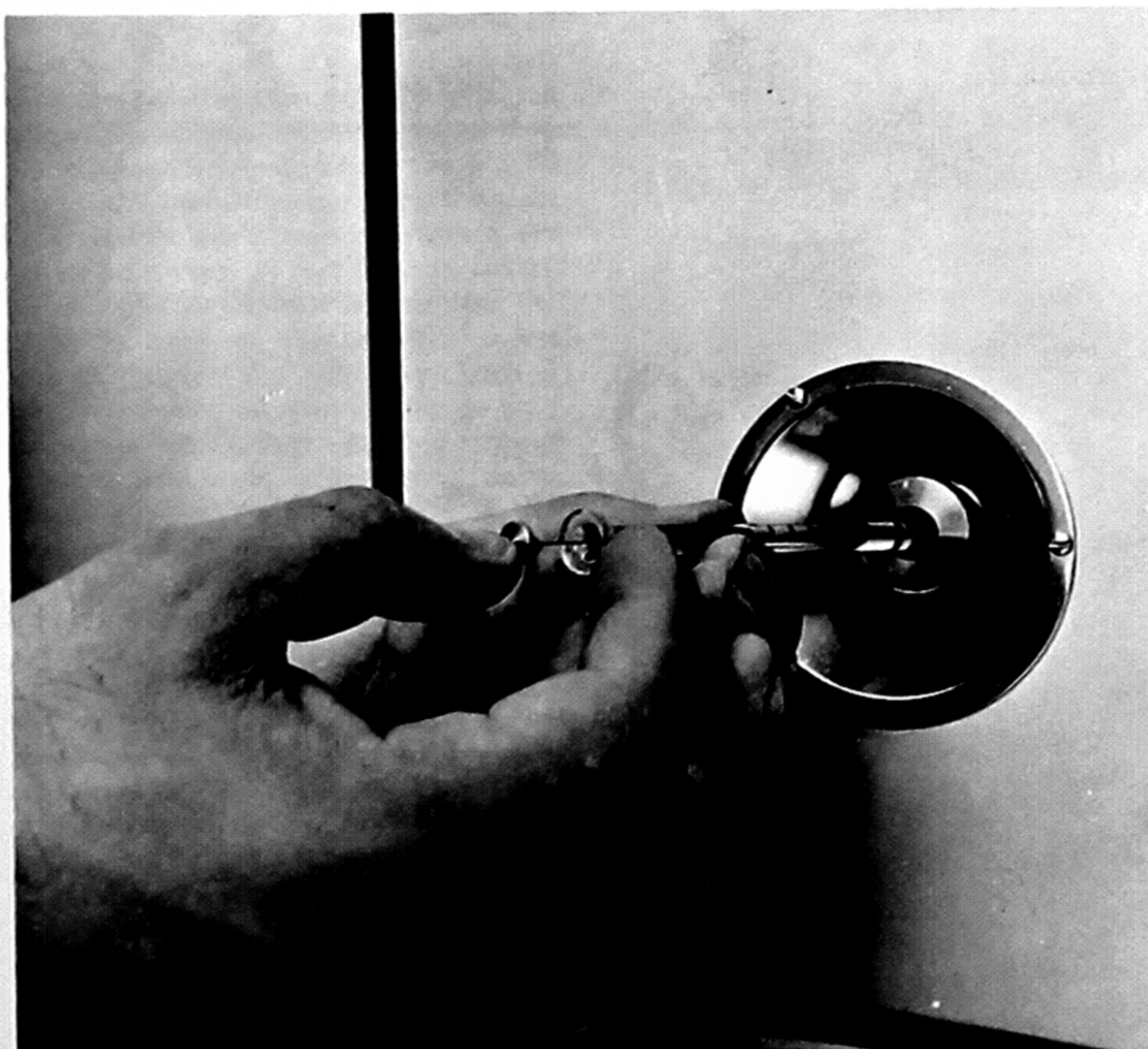
The most pervasive problem in chemistry and biochemistry is that of determining the composition of complex mixtures of matter. The research chemist would like to know at all times exactly what his test tubes and flasks contain; the industrial chemist has an equally urgent need to know the composition of materials flowing through reaction vessels and distillation columns

and into product tanks. Within the past 20 to 30 years, in response to this need for fast and accurate analyses, a number of powerful analytical tools have been developed. They include instruments that measure how compounds absorb ultraviolet and infrared radiation, instruments that determine atomic or molecular mass and instruments that determine how the magnetic properties of

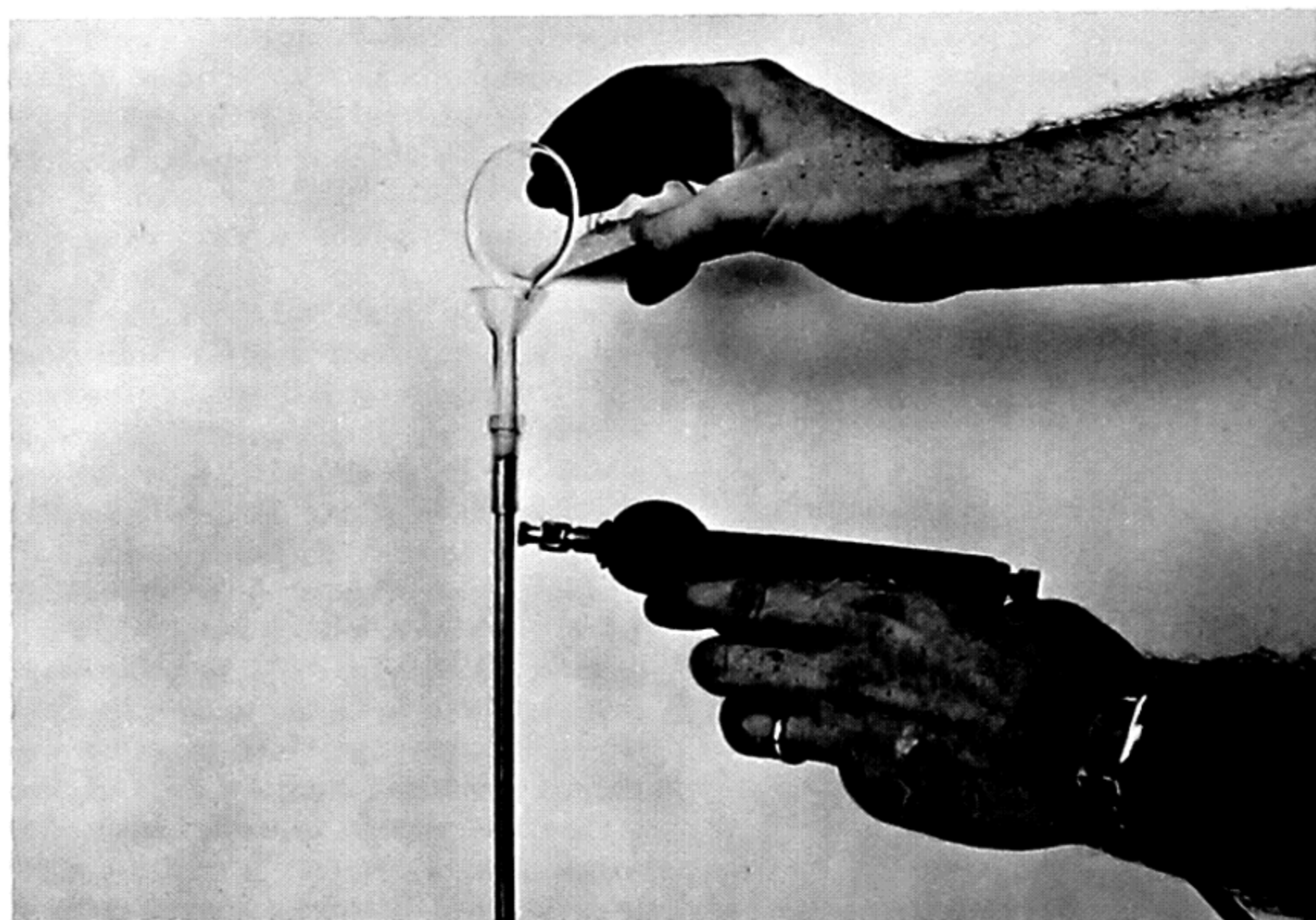
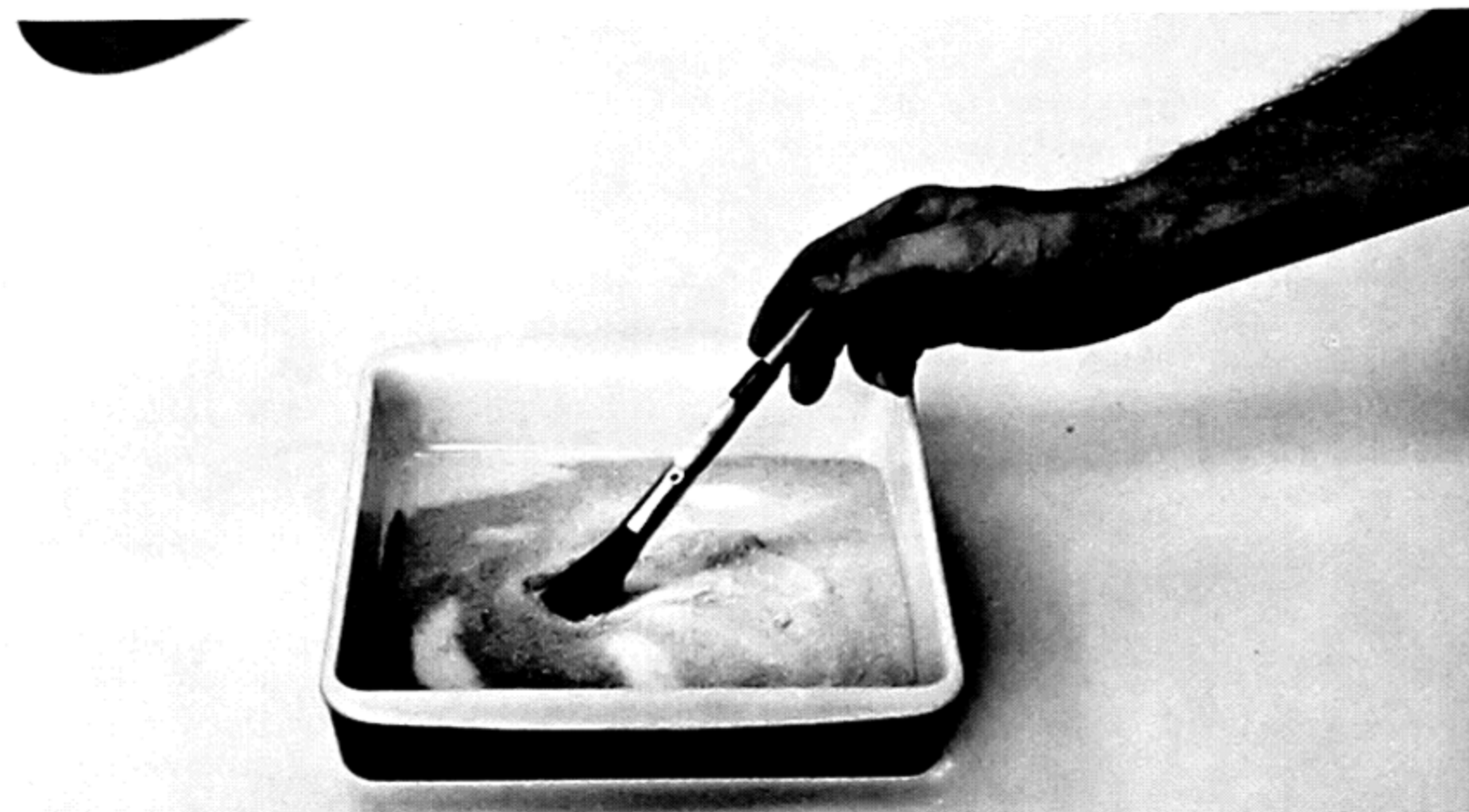
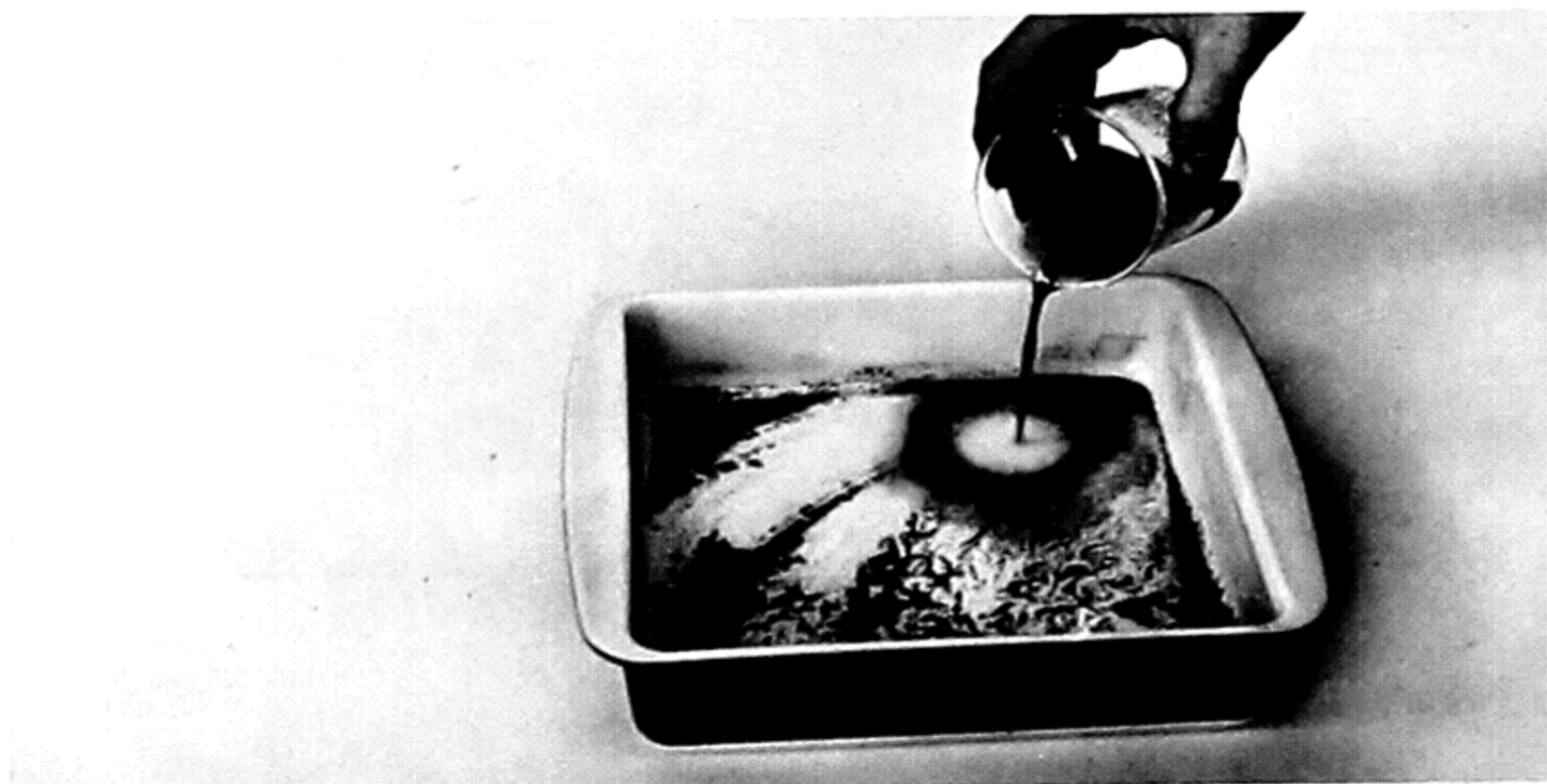
atomic nuclei are influenced by different molecular configurations.

One of the newest and most versatile analytical techniques is gas-liquid chromatography, usually referred to simply as gas chromatography. The name is somewhat misleading; there is nothing chromatic about the method or its results. The name comes from the original method of liquid-solid chromatography described in 1906 by Michael Tswett, a Russian botanist. Tswett found that when a solution of chlorophyll was allowed to filter through a column firmly packed with pulverized calcium carbonate, the various fractions of the chlorophyll mixture separated into distinctively colored bands. He called the result a chromatogram [see "Chromatography," by William H. Stein and Stanford Moore; SCIENTIFIC AMERICAN Offprint 81]. A useful variation of Tswett's original concept is paper chromatography, in which compounds in solution migrate at different speeds across a sheet of porous paper.

In gas chromatography the compounds in a mixture migrate at differing speeds when carried along by an inert gas through a tube that has been packed or treated in a special way. The method was first suggested in 1941 by the British chemists A. T. James and A. J. P. Martin. The method underwent development in many laboratories beginning around 1950, and the first commercial instrument came on the market in 1955. Today some 40 models of gas chromatographs are being built by 20 U.S. and a dozen European manufacturers. The instruments are widely used for analyzing complex organic mixtures such as those commonly found in petroleum products, essential oils, perfumes, flavors and other substances of biological origin. Samples containing as many as 76 dif-



SAMPLE IS INJECTED into a gas chromatograph by means of a syringe that can deliver liquid samples as small as one hundred-thousandth of a cubic centimeter. Gas samples require a special valve arrangement to meter the sample through a chamber of known volume.



PREPARATION OF CHROMATOGRAPHIC COLUMN begins (*top*) with pulverized diatomaceous earth, to which is added a viscous partitioning liquid dissolved in a volatile carrier. The solvent is evaporated by heat while the coated powder is gently agitated (*middle*). The dried coated powder is packed into a tube with the aid of a vibrator (*bottom*).

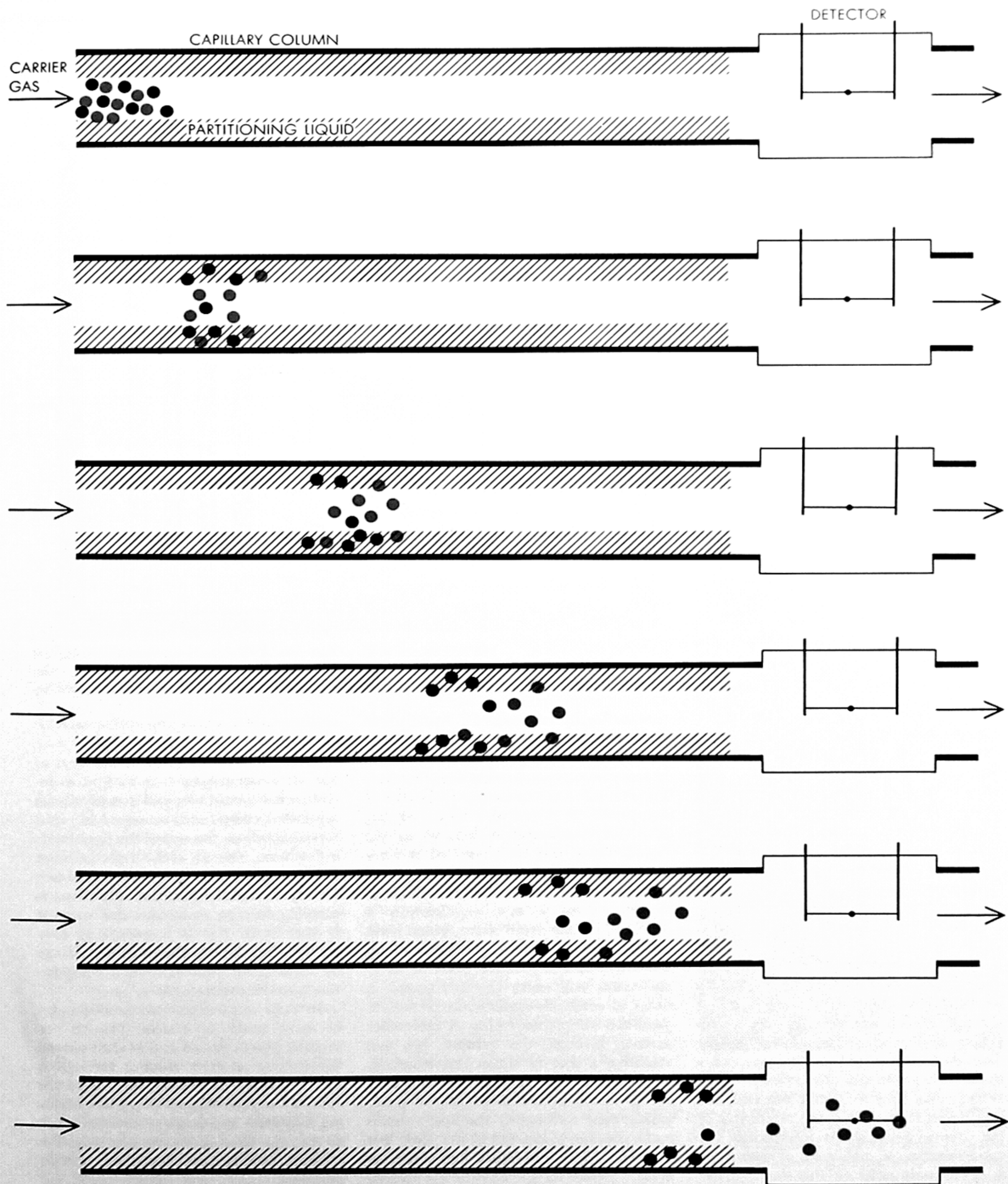
ferent substances have been successfully analyzed in one pass. Analysis time is typically a few minutes and sometimes only a few seconds. Samples usually range in size from a few hundredths to a few thousandths of a gram. Some instruments can handle samples weighing not much more than a millionth of a gram, and in such samples they can detect the presence of substances that weigh no more than a trillionth of a gram—about the weight of a single bacterium.

The fractionating column of a typical gas chromatograph consists of a copper or stainless steel tube about a quarter of an inch in diameter and from one to four meters long, though occasionally much longer tubes are used. The tube is packed with an inert material such as firebrick or diatomaceous earth that has been pulverized and coated with a non-volatile liquid called a partitioner [*see illustration at left*]. After the tube has been packed it is usually bent into a series of U turns or wound into a helix so that it can be fitted easily into an insulated box, the temperature of which is thermostatically controlled.

The liquid selected as the partitioner largely determines the performance of the chromatograph. The partitioner must not react with the sample being analyzed and it must not be volatilized by the stream of carrier gas that propels the sample through the column. Above all, the partitioner must show different affinities (to use an old-fashioned term) for each of the substances likely to be found in the sample mixture.

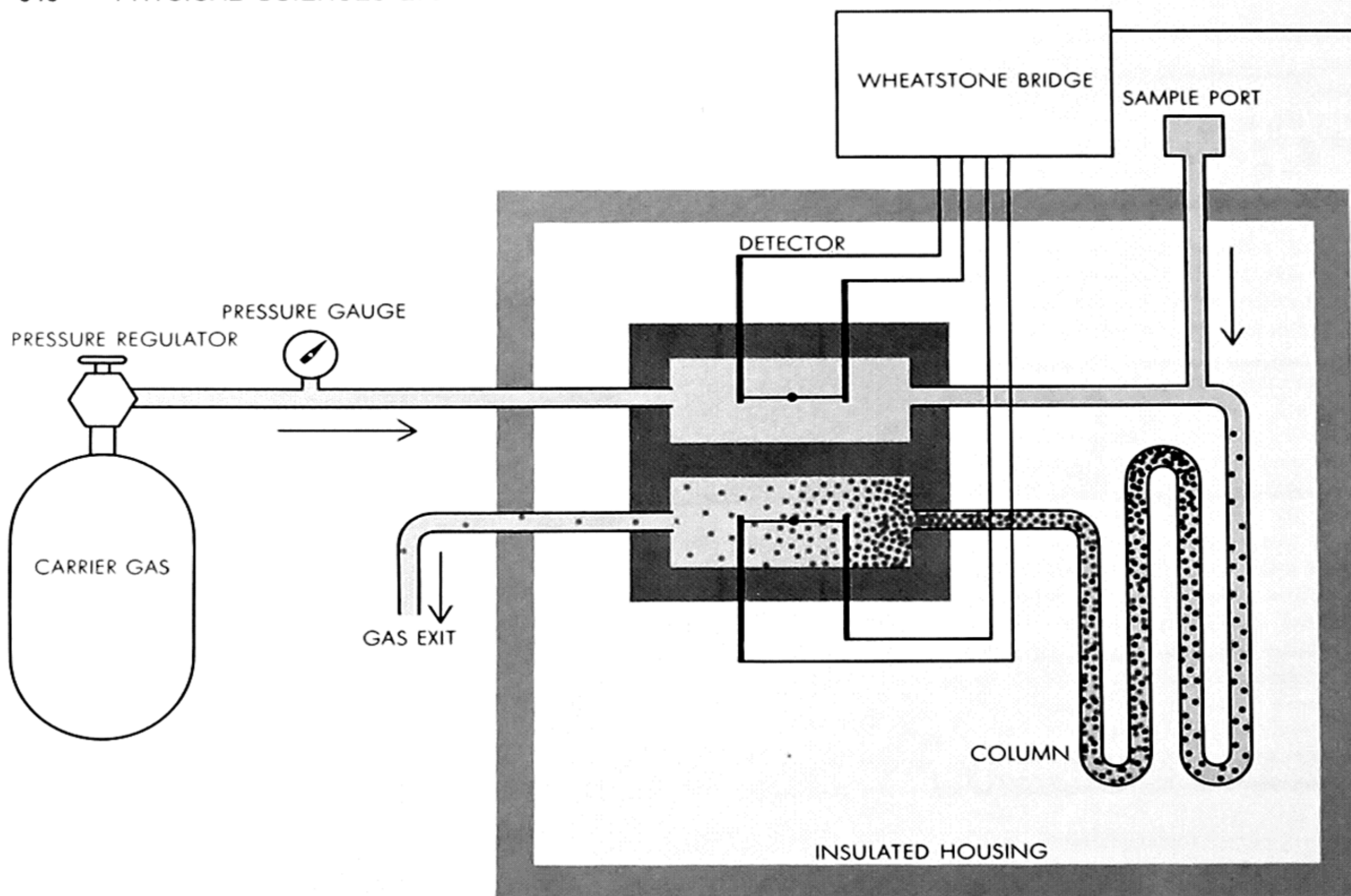
Partitioners that work well with one type of sample may be completely useless for another type. As the sample is moved through the column by the carrier gas, the partitioner must interfere in a selective fashion with the progress of each compound present, slowing up the progress of some and letting others travel through the column more swiftly. At the outlet of the column a detecting device signals the emergence of each different compound by activating a recording pen on a strip chart.

In the search for good partitioners, builders of gas chromatographs have experimented with virtually every viscous fluid, grease and low-melting-point solid in the laboratory, including such substances as silicone rubber, stopcock grease and hydrogenated shark oil. This eclecticism has sometimes had unhappy results when excellent separations were achieved with substances that could never be duplicated, even when re-



CHROMATOGRAPHIC SEPARATION takes place when a sample mixture (*black and colored balls*) is driven by an inert gas through a capillary tube coated with an immobile liquid called a partitioner. In the original form of gas-liquid chromatography the partitioner

was deposited on a pulverized packing. The role of the partitioner is to dissolve (and adsorb) various components of the sample in differential fashion. After fractionation the separated components pass through a detector, whose output is recorded on a chart.



SIMPLICITY OF GAS CHROMATOGRAPH is among the virtues of the instrument. The sample is swept by a carrier gas through a specially treated column in which various components of the sam-

ple migrate at different speeds. A detector measures the electrical conductivity of the gas leaving the column as well as that of the carrier gas entering the column. The difference, as determined by

ordered by the same lot or batch number. In order to deposit the viscous partitioner on the pulverized support, the partitioner is usually dissolved in a volatile solvent that can be evaporated, leaving the partitioner behind. Normally the partitioner weighs from about a quarter to a third as much as the pulverized packing in the column. The coated particles should flow freely, and there is some advantage in holding the amount of partitioner to a minimum.

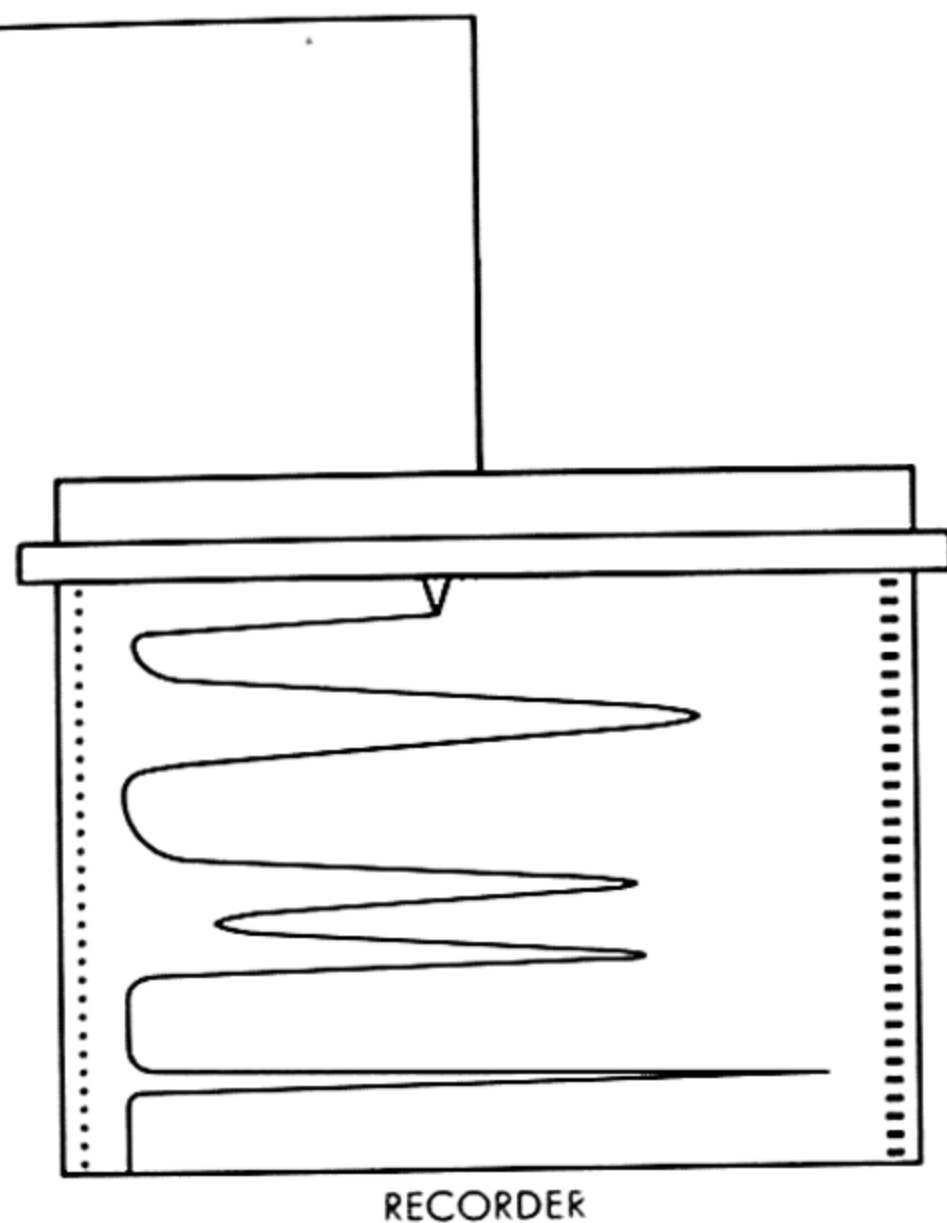
Let us now look more closely at what happens to a collection of molecules injected into the carrier gas moving through the column of a gas chromatograph. Some of the molecules rapidly dissolve in the liquid partitioner, and a dynamic equilibrium is soon established as they pass back and forth between the liquid and vapor filling the interstices of the column packing. At equilibrium the concentration of molecules of each type is a constant ratio in the two phases. Molecules of compound A, for example, may partition themselves equally between the liquid and vapor phase; molecules of compound B, on the other hand,

may be highly soluble in the liquid phase and therefore relatively few of them will be found in the vapor phase once equilibrium is reached. In this case the moving gas will tend to sweep molecules of compound A down the tube, leaving those of compound B behind in the liquid. Once the molecules of A have been carried to a region containing fresh liquid, however, some of them will redissolve until a new equilibrium is reached. By the same token, when fresh gas passes over the liquid containing molecules of B in solution, some of the B molecules will enter the gas phase in order to establish equilibrium. If we regard the sample as a plug of molecules moving through the column, we can visualize a sort of molecular leapfrog. The volatile molecules are continuously being swept to the head of the plug, where they redissolve; the less volatile molecules fall to the tail of the plug, but they too are continuously being picked up and inched forward by the gas stream pressing from behind. Eventually, if conditions are right, all the molecules of the more volatile components will be carried clear ahead of those of the less

volatile and a clean separation will be achieved.

A major problem in the early days of gas chromatography was to find a detector that would respond quickly as the separated compounds emerged in rapid succession from the end of the fractionating column. The job of the detector is not to identify the emerging compound but merely to signal when the output gas is carrying foreign molecules and when it is not. Once this is known it is easy enough to calibrate the output readings by feeding samples of known composition into the instrument.

In their original gas chromatography of fatty acids or amines (which are organic bases) James and Martin passed the output of their column through a solvent that extracted the acids or amines from the effluent gas. The collecting was done in a series of small batches so that the time of emergence from the column could be recorded. Using a color indicator, James and Martin could tell when an acid or an amine emerged from the column. They then titrated each sample to determine how much acid or base was present. The job not only was



a Wheatstone bridge, is recorded on a strip chart. The instrument can be calibrated by analyzing samples of known composition.

laborious but also it lent itself to titratable samples only.

A much simpler and more universal solution to the detection problem was finally found in the thermal-conductivity cell. The cell utilizes the fact that the electrical resistance of a heated wire varies with its temperature. If a gas of constant composition and flow rate is allowed to pass over a heated wire, the wire will be cooled a constant amount and so register a constant resistance. If a gas of different thermal conductivity appears in the stream striking the wire, the wire will change in temperature and in electrical resistance, and this change can be recorded in ink on a strip chart.

At least two other detection devices have been developed for gas chromatography. In both devices a change in gas composition is signaled by a change in the ionization—and hence the electrical conductivity—of the gas stream. In one device the gas stream is ionized (broken up into electrically charged fragments) by being passed through a hydrogen flame; in the other device the stream is ionized by bombardment with radiation from a bit of radioactive material.

To obtain sharply defined gas-chromatograph records, called fractograms, the instrument designer can vary the pressure and flow rate of the carrier gas, the operating temperature of the column, the structure and particle size of the column packing and, of course, the nature of the liquid partitioner. A considerable body of theory and empirical art has grown up around the solute-solvent interactions that underlie effective partitioning.

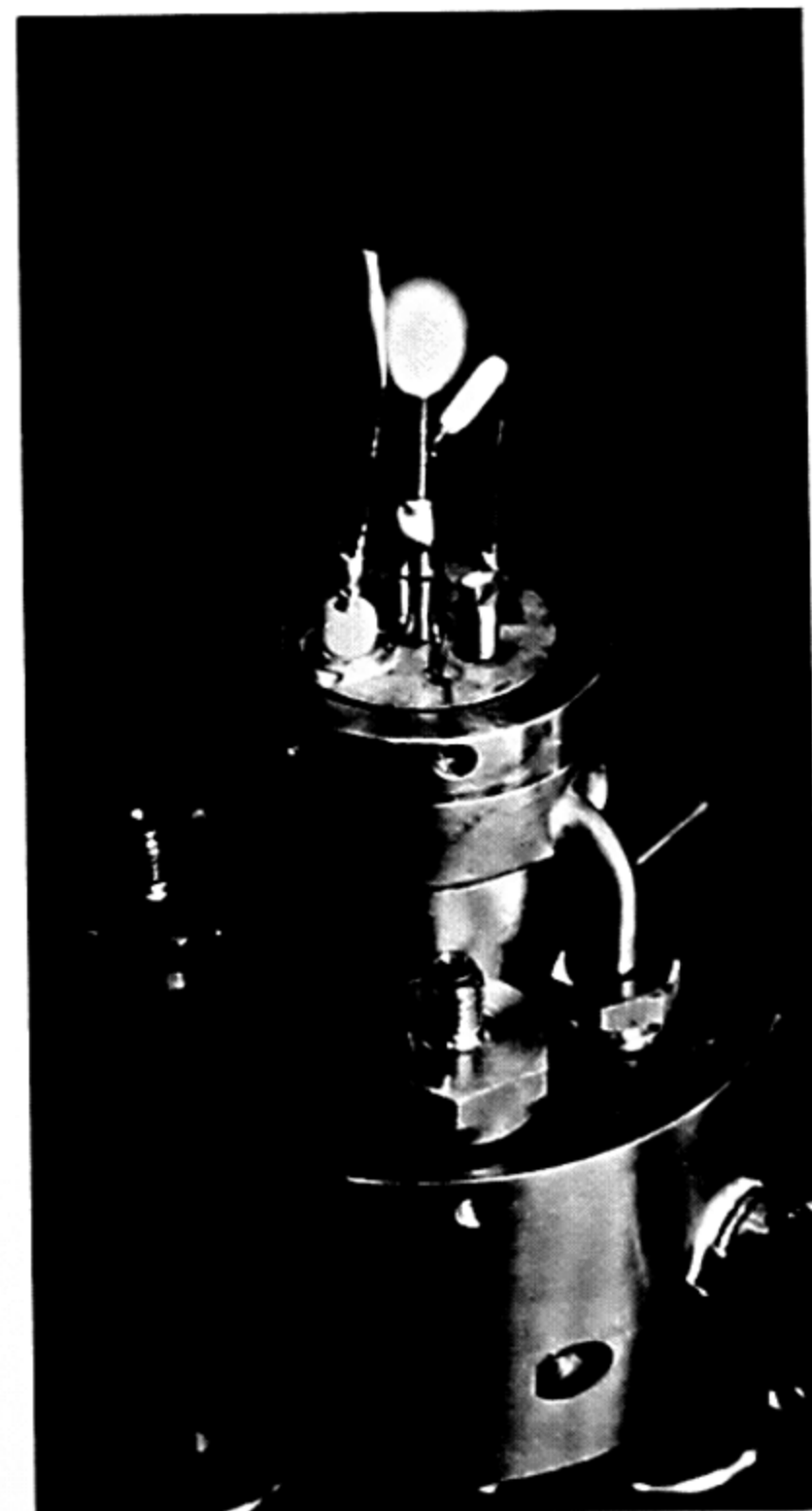
For example, if one wishes to separate a hydrocarbon and an alcohol having nearly the same boiling point (*e.g.*, 3-methylpentane and methyl alcohol, both of which boil at about 64 degrees centigrade), it is desirable to use a partitioner that resembles the alcohol in containing hydroxyl (OH) groups in its structure. The hydrogen atom of the hydroxyl group in the alcohol will tend to form a bond with the oxygen atom of the hydroxyl group in the partitioner. The hydrocarbon will not form such a bond and will therefore move through the column faster. A relatively nonvolatile liquid containing hydroxyl groups that can be used for this separation is polyethylene glycol.

Offhand one might think that a partitioner, to be effective, should show a differential solvent action on each component in a sample mixture. One might then conclude that the solutes would be retained by a partitioner in order of their solubility. In practice, however, this is not always true. The reason is that solubility as conventionally measured with solvents in bulk—say in a laboratory beaker—is quite different from the solubility shown when a solvent is thinly distributed over an enormous surface area, as it is in a chromatographic column. In the latter case a new factor appears: the effect of surface adsorption. A solute is said to be adsorbed if its concentration in the immediate region of the surface exceeds that in the bulk liquid. Adsorption can arise either at the interface between the solid support and the partitioning liquid, or at the interface between liquid and gas, or it can occur in both regions. Sometimes adsorption enhances the desired separation; at other times it interferes with it. For example, when alcohols are chromatographed on a hydrocarbon partitioner, they tend to displace the hydrocarbon and fasten themselves to the support.

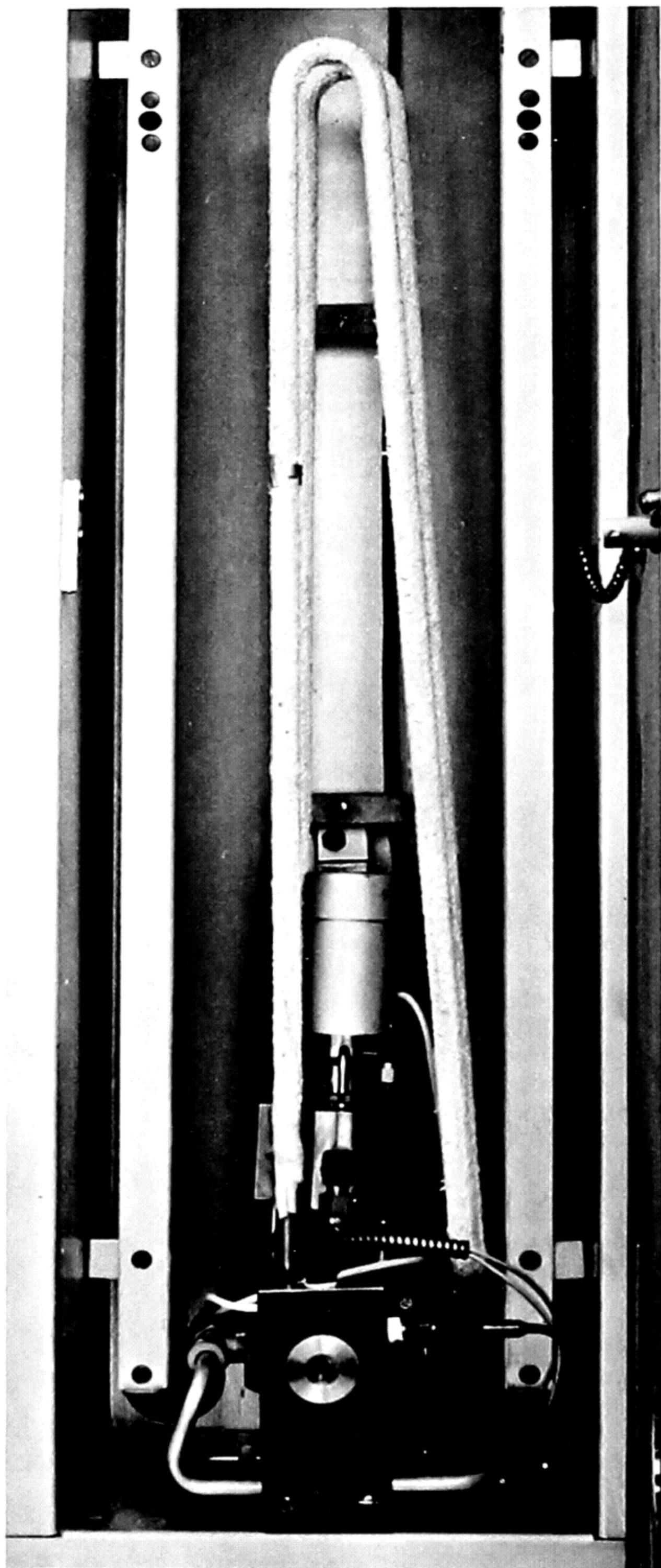
Considerable work has been done on gas chromatographs that achieve separations strictly by differential adsorption on the solid packing, without help from a liquid partitioner. This is called gas-solid chromatography, and for certain

sample mixtures in which the molecules have much the same architecture it produces even sharper separations than gas-liquid chromatography.

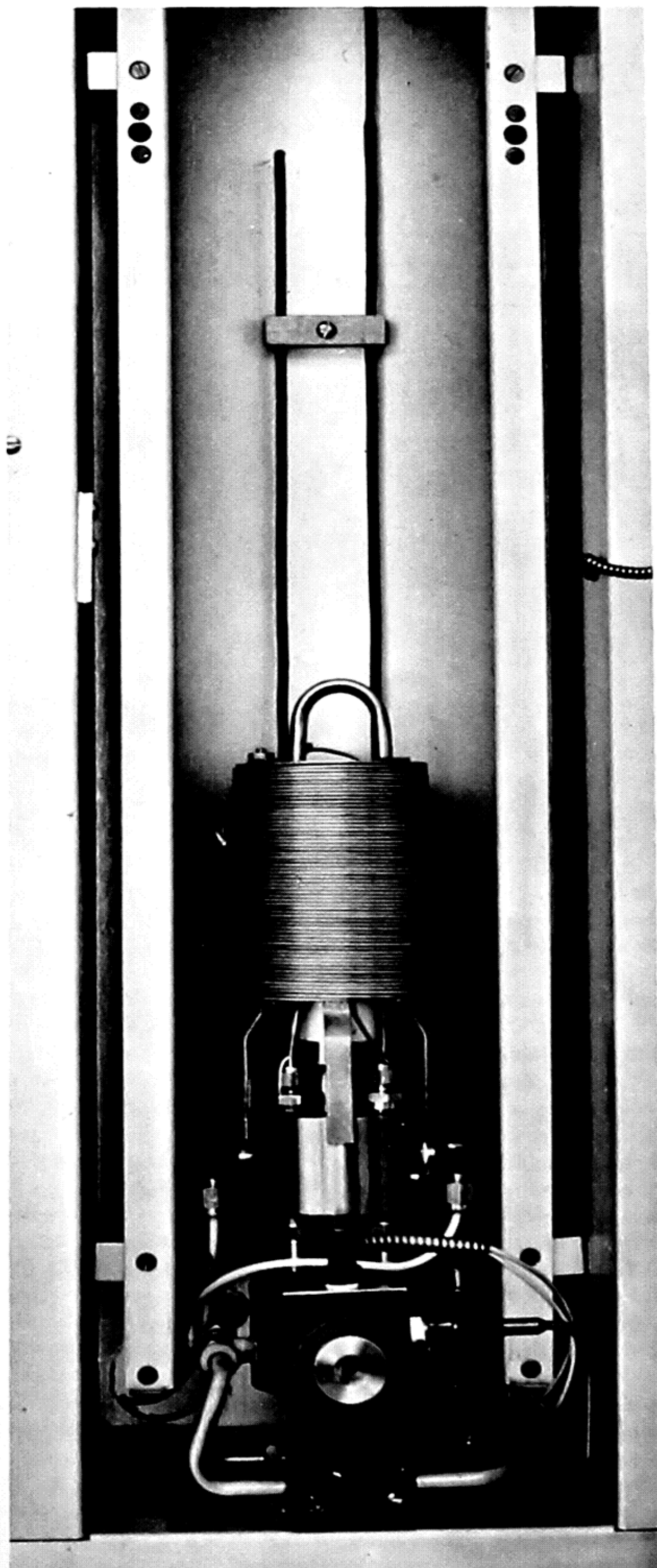
Until recently gas-liquid chromatography was successful only with sample mixtures whose components boiled within about 50 degrees C. of one another. When the boiling range was greater than that, it was usually impossible to choose a column operating temperature that would sharply resolve both the least volatile and most volatile components. If the temperature was held low, the more volatile substances would be resolved, but the less volatile would lag behind and become spread out by diffusion. By raising the temperature the resolution of the less volatile could be sharpened, but the more volatile would then rush through the column and emerge in a poorly resolved bunch. The



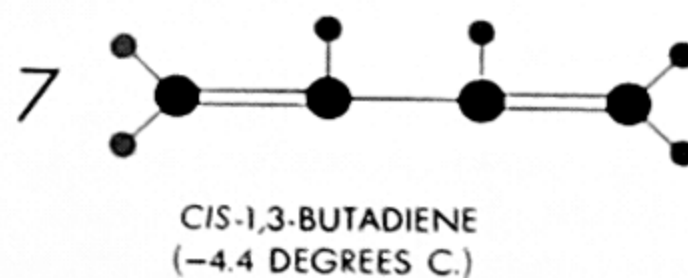
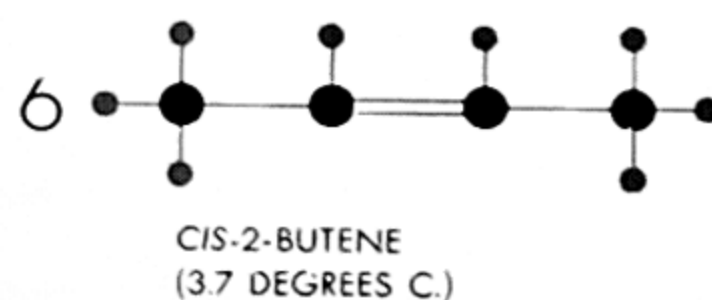
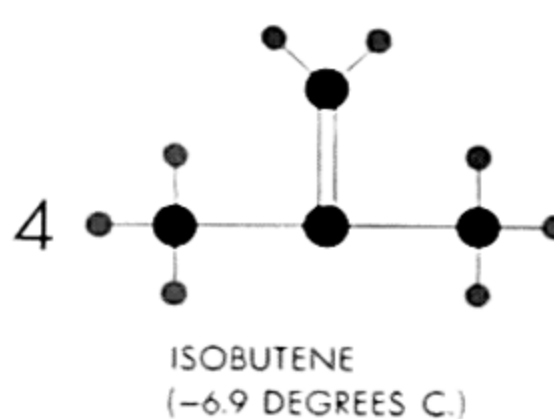
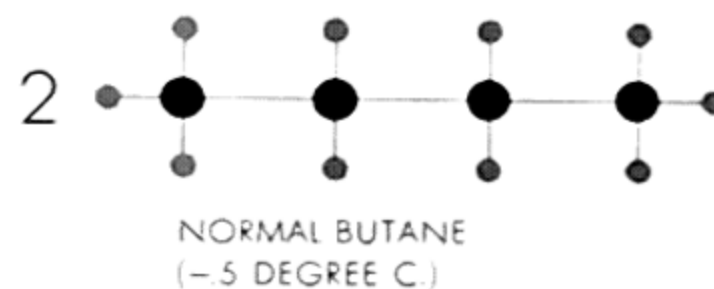
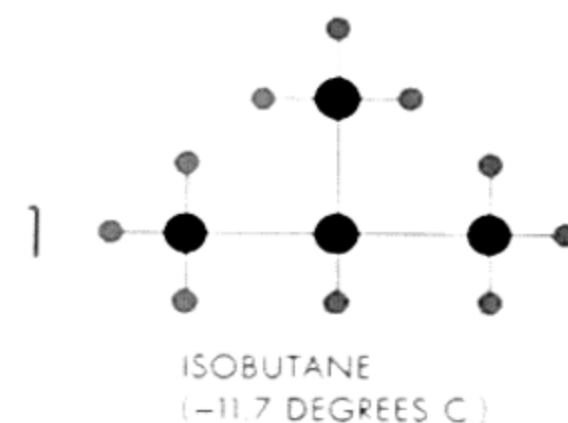
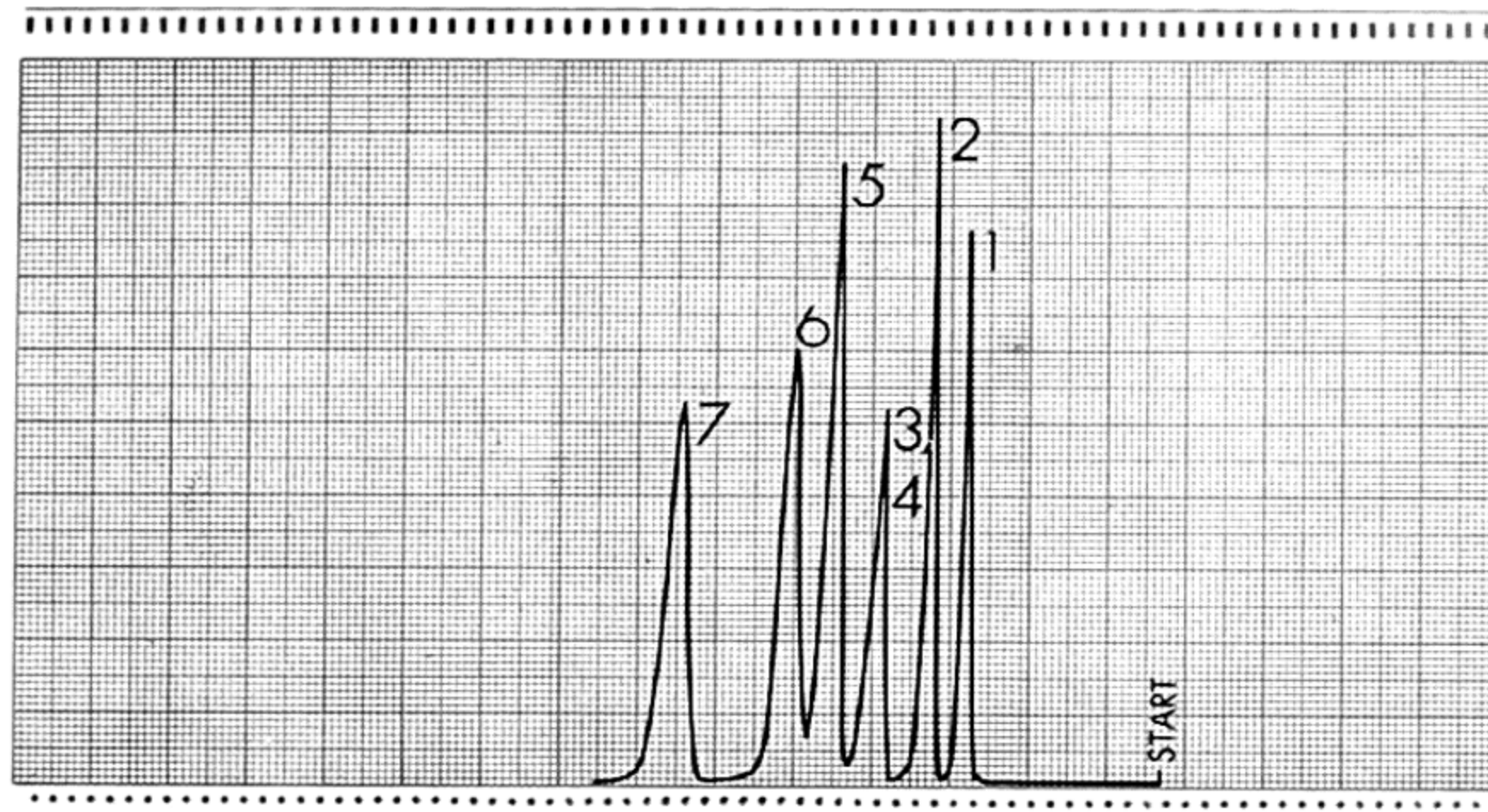
IONIZATION DETECTOR, one of three principal types of detector used in gas chromatography, employs a hydrogen flame to break up chemical compounds into electrically charged fragments (ions). By measuring the electrical conductivity of the ionized gas at the column exit the detector signals the passage of various fractions of the sample. In this photograph the cover of the detector has been removed and the normally colorless flame has been made visible.



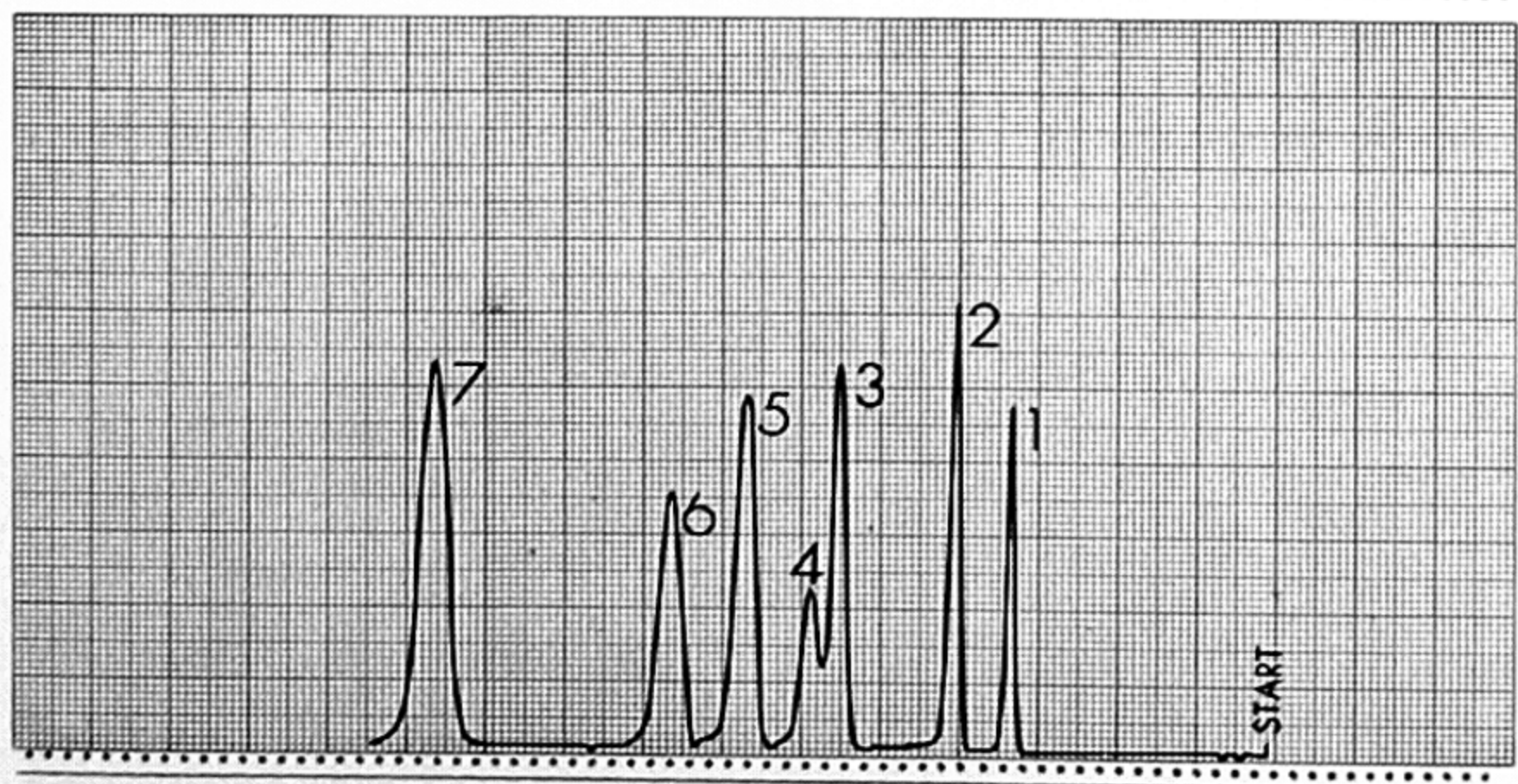
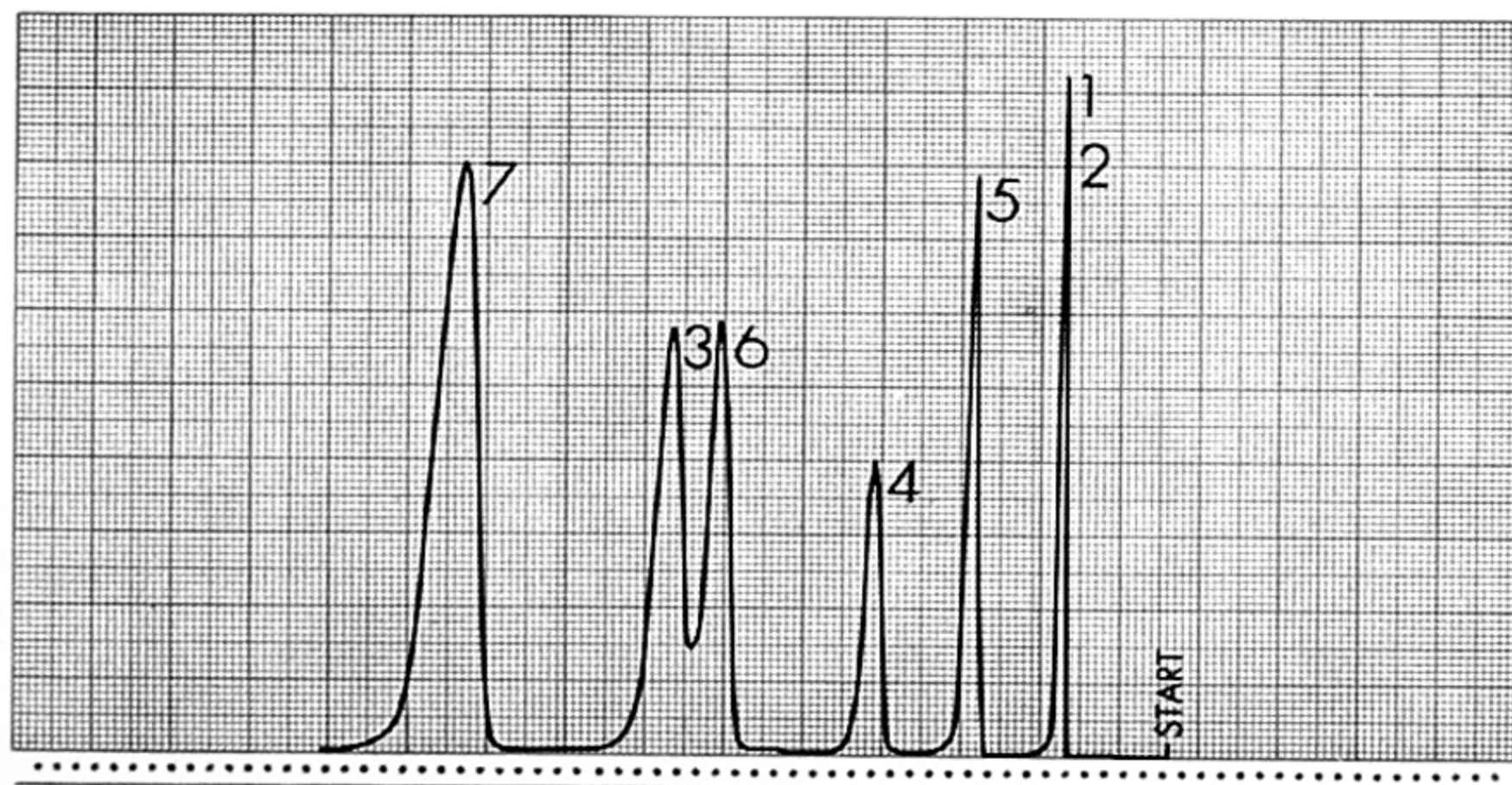
TWO TYPES OF GAS CHROMATOGRAPH differ in the design of the column where fractionation takes place. The instrument at left uses the original type of column, a quarter-inch in diameter and one to several meters long, packed with a pulverized inert substance. The column illustrated is four meters long and is folded



twice. Instrument at right uses a capillary column, 150 to 300 feet long, without packing, proposed by Marcel J. E. Golay, a consultant to the Perkin-Elmer Corporation. A nonvolatile liquid carried on the packing, or on the inside wall of the capillary tube, acts as a partitioning agent and promotes fractionation of the sample.



● CARBON ● HYDROGEN



SEPARATION OF FOUR-CARBON HYDROCARBONS, which boil within a narrow range, is a familiar problem in the oil and synthetic rubber industries. The molecular structure and boiling points of the principal four-carbon hydrocarbons appear at right. The fractograms (*left*) show how these compounds are fraction-

ated by different partitioning liquids: dimethylsulfolane (*top*), silver nitrate in diethylene glycol (*middle*) and hexanedione (*bottom*). The upper two fractograms were made on columns four meters long operated at 25 degrees centigrade. Bottom fractogram was made on a column two meters long operated at zero degrees.

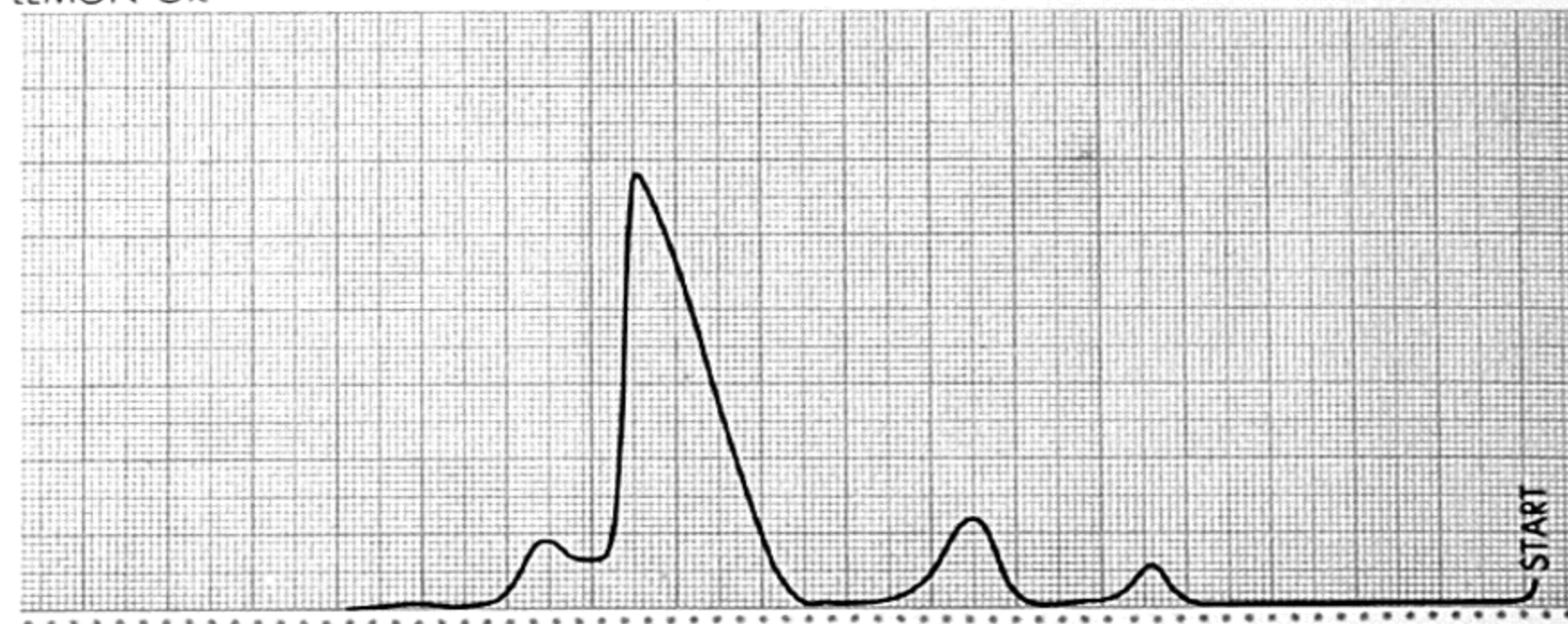
answer to this problem has been found in "temperature programming," which simply means starting the separation at a low temperature and raising it in regular steps until the job is done. By this procedure one can fractionate mixtures whose components have a boiling range as broad as 200 degrees C.

Actually it is not necessary for the components of a sample to be anywhere near their boiling points for gas chromatography to work. At the University of Arizona we have analyzed samples of volatile inorganic halides (for example, niobium chloride) at temperatures near their melting point, or in the vicinity of 250 degrees C. Other investigators have operated machines at 500 degrees C. that will chromatograph compounds boiling as high as 625 degrees C.

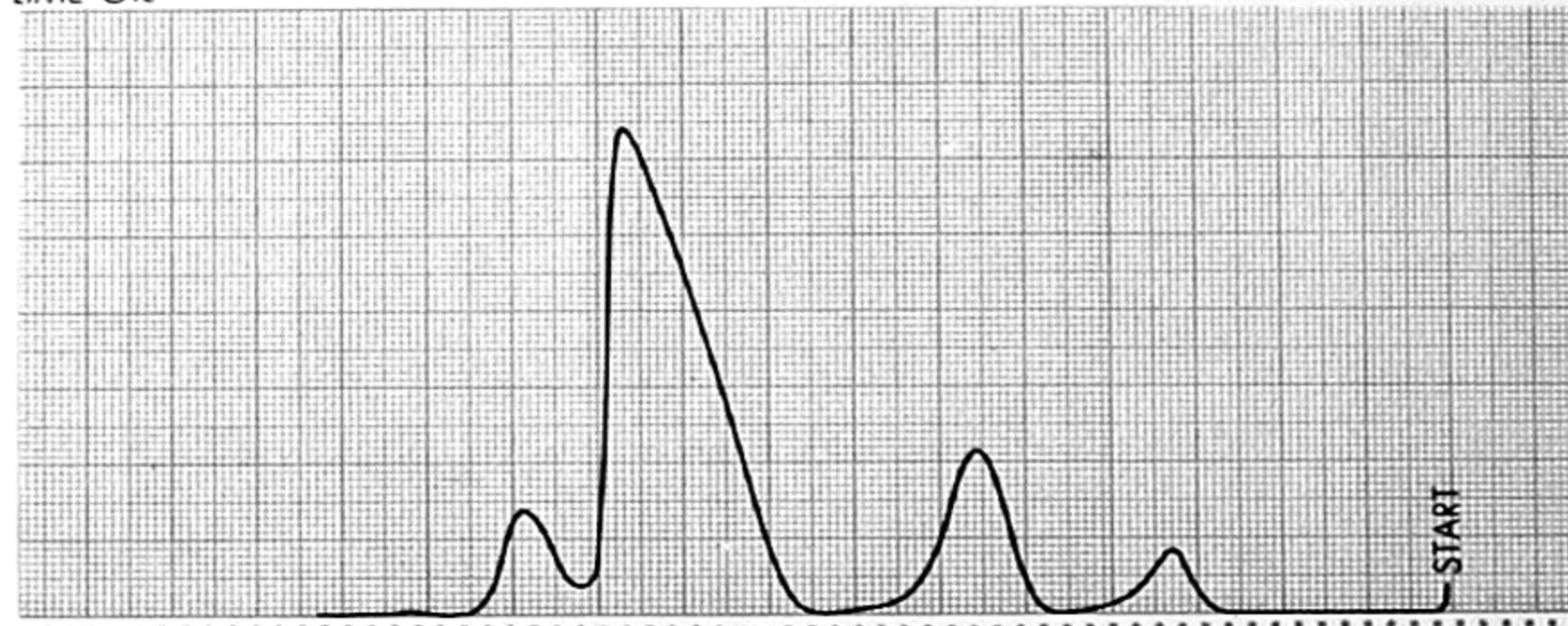
An important advance in the design of gas chromatographs was made about 1956 by Marcel J. E. Golay, a consultant to the Perkin-Elmer Corporation. Golay, a student of information theory, conceived the improvement after making a theoretical study of the migration of a solute through a packed column. In the fashion characteristic of theorists he sought a simplified model to substitute for the actual highly complex situation prevailing in a tube packed with porous particles of random shapes and sizes. He selected for his model a bundle of capillary tubes, equal in diameter to the granule size of the packing, which were evenly coated with the partitioning medium. Upon analyzing his calculations Golay concluded that a single, very long coated capillary tube would achieve separations equivalent, if not superior, to those produced by packed columns and do so in a much shorter time and under less severe conditions of temperature and driving pressure. Golay met his skeptics by preparing some of the first columns himself. They amply fulfilled his predictions.

Commercial columns of his design usually employ a capillary tube having an inside diameter of .01 inch. The tubes, ranging from 150 to 1,000 feet in length, are coiled into a compact helix [see illustration at right on page 648]. If these capillary columns are not to be flooded, the sample size must be extremely small, usually about five-millionths of a cubic centimeter in liquid volume, or about five-thousandths of a cubic centimeter after vaporization. To obtain such minute volumes a sample some 20 times larger is vaporized and shot through a stream splitter that ad-

LEMON OIL



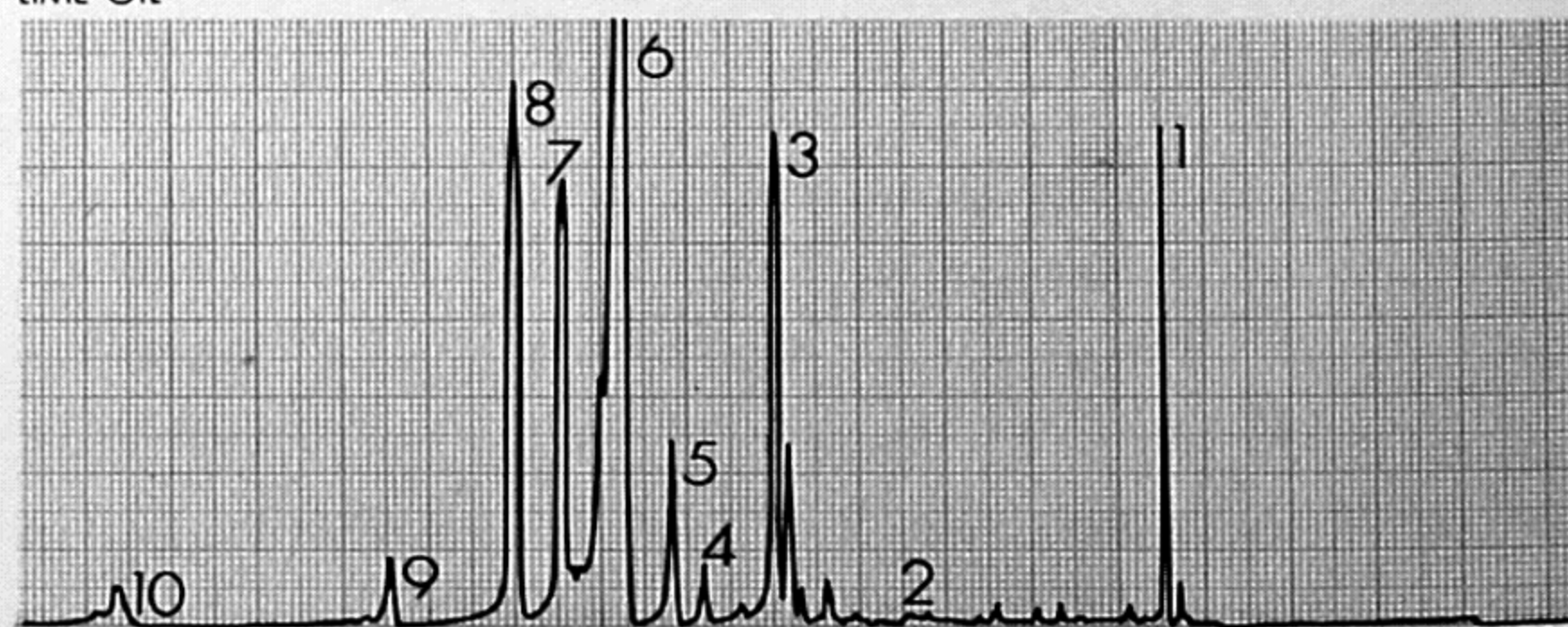
LIME OIL



LEMON OIL

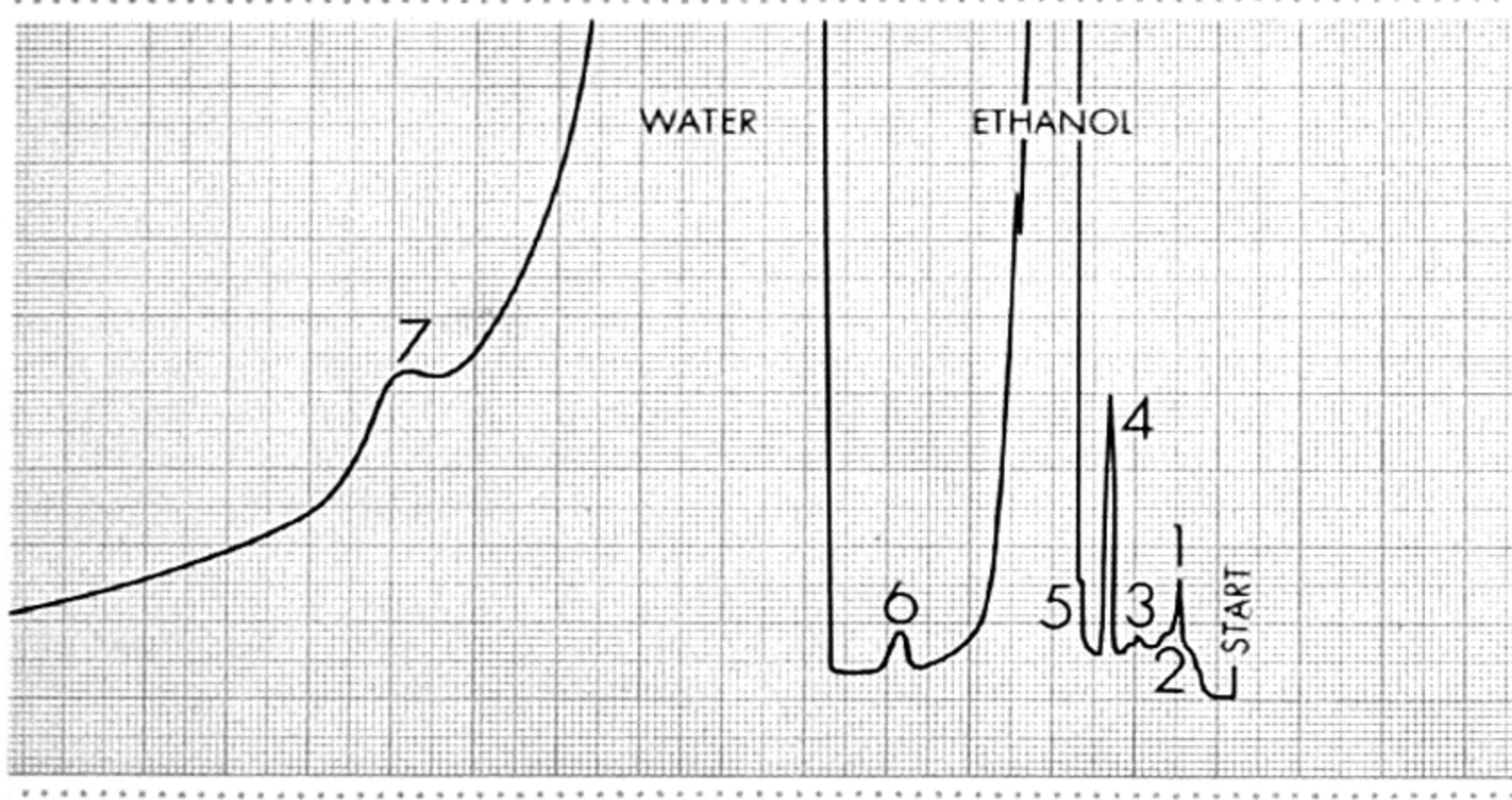


LIME OIL



PERFORMANCE OF OLD AND NEW INSTRUMENTS is demonstrated by fractograms of two similar essential oils: the terpene fractions of lemon oil and lime oil. When analyzed on a standard two-meter packed column, the two oils produced the upper pair of fractograms. When analyzed on a 175-foot capillary (Golay) column, they produced the detailed lower pair. The peaks labeled "6" are made by limonene ($C_{10}H_{16}$), which has a lemon-like odor.

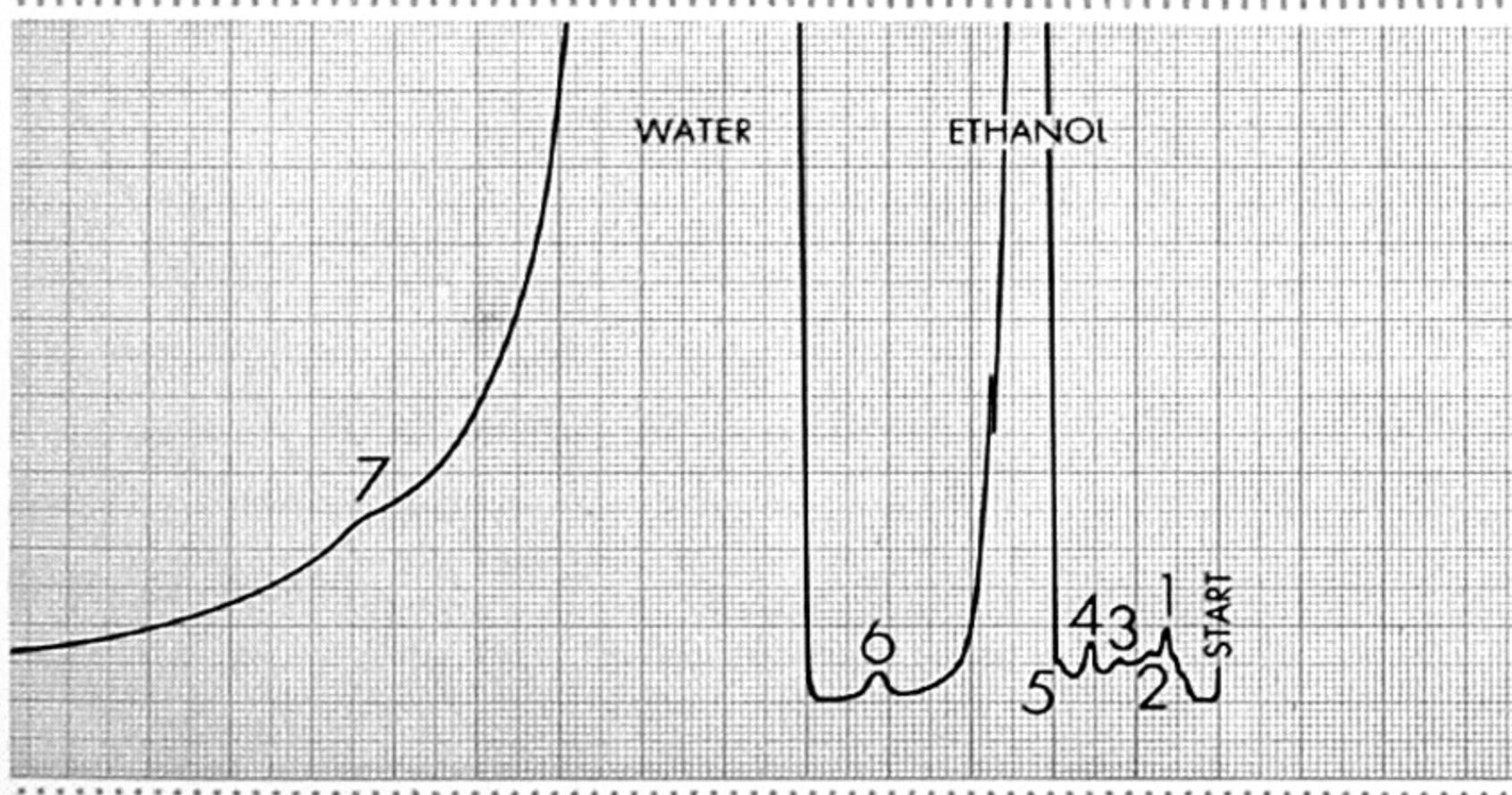
BOURBON



- 1 ACETALDEHYDE
- 2 FORMALDEHYDE
- 3 ETHYL FORMATE
- 4 ETHYL ACETATE
- 5 METHANOL
- 6 PROPANOL
- 7 ISOAMYL ALCOHOL

mits about 5 per cent to the column and discards the remainder. Ordinarily a few minutes to half an hour are required for such samples to migrate through the column. As an extreme example of what can be achieved with a capillary column and a highly sensitive detector, it has been possible to record 14 peaks—each representing a separate compound in a sample of closely related hydrocarbons—in less than two seconds.

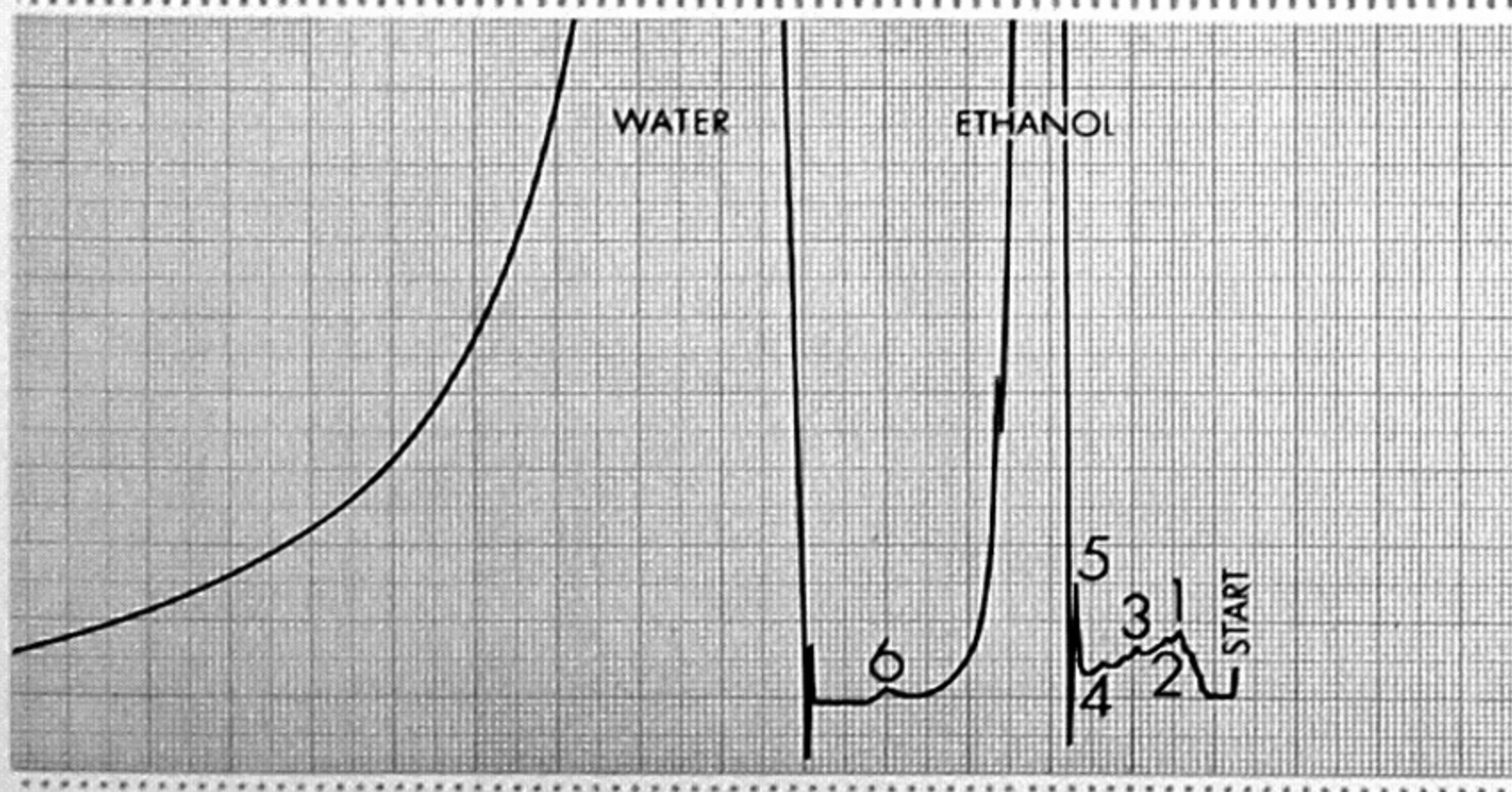
SCOTCH



The areas under different peaks in a fractogram are roughly proportional to the fractional amounts of each substance in the original sample. With care the method is accurate to about 2 per cent. In general the investigator has two methods for discovering exactly what substance is represented by a particular peak. The commonest method is to use samples of known composition to calibrate the machine. Alternatively he may isolate the column effluent that produced a given peak and characterize it by some suitable analytical technique, for example by using an infrared spectrophotometer or a mass spectrometer. The gas chromatograph will separate not only compounds with closely similar properties but also various forms of individual compounds. In organic chemistry almost all but the simplest compounds can exist in two or more forms known as isomers. These are molecules containing the same number and kind of atoms fitted together in different geometrical arrangements [see illustration on page 649].

In some cases gas chromatography provides direct clues to compound identification. The volume of gas, called the retention volume, that precedes a particular solute through the column depends on the nature of the solute, the choice of partitioner and the temperature. Within limits, retention volume is not overly sensitive to length of column, driving pressure, flow rate or the amount of partitioner employed. As a result one can determine retention volumes for various compounds of interest and use these volumes for identifying unknown samples [see illustration on following page]. A considerable effort is now being made to utilize gas chromatography for determining the structure of molecules by

GIN



FRACTOGRAMS OF POTABLE SPIRITS provide a sensitive measure of volatile components that influence taste and aroma. These analyses confirm that Scotch and gin contain fewer such components than bourbon. (Isoamyl alcohol is fusel oil.) Samples were run on a two-meter packed column. Peaks were identified by adding various known compounds to a whiskey that had previously been analyzed. Analyses were made by Robert B. Carroll, a chemical consultant, and Lawrence C. O'Brien, then at the Perkin-Elmer Corporation.

relating retention volume to particular molecular configurations.

Much of the explosive growth of gas chromatography can be attributed to the petroleum industry, which has to deal with materials of extraordinary complexity. Crude oils commonly contain more than 150 different hydrocarbons, many of them isomers of each other. For separating them gas chromatography has proved a powerful tool. Oil firms have now begun to use gas chromatographs for continuous monitoring of process streams in the refinery. Chromatographic analyses will be transmitted to a computer, which will automatically calculate the optimum operating conditions for catalytic cracking towers. Gas chromatography is almost the only method available for detecting certain catalytic poisons that impair the polymerization process when present in concentrations of only 50 parts per million.

Automotive engineers use gas chromatography for analyzing the exhaust of engines under a variety of operating conditions with various fuels. The results are used to improve both engines and fuels and also contribute to a reduction in the air pollution created by engine exhausts.

Elsewhere, fractograms are rapidly replacing the opinions of "sniff and taste" panels as a method of assaying the uniformity of instant coffees, blends of whiskey and many other products whose commercial acceptance depends on subtle flavors and aromas. In such products gas chromatography can often detect trace components present in only one part per billion. It is, in fact, the first analytical instrument to rival the human nose in sensitivity.

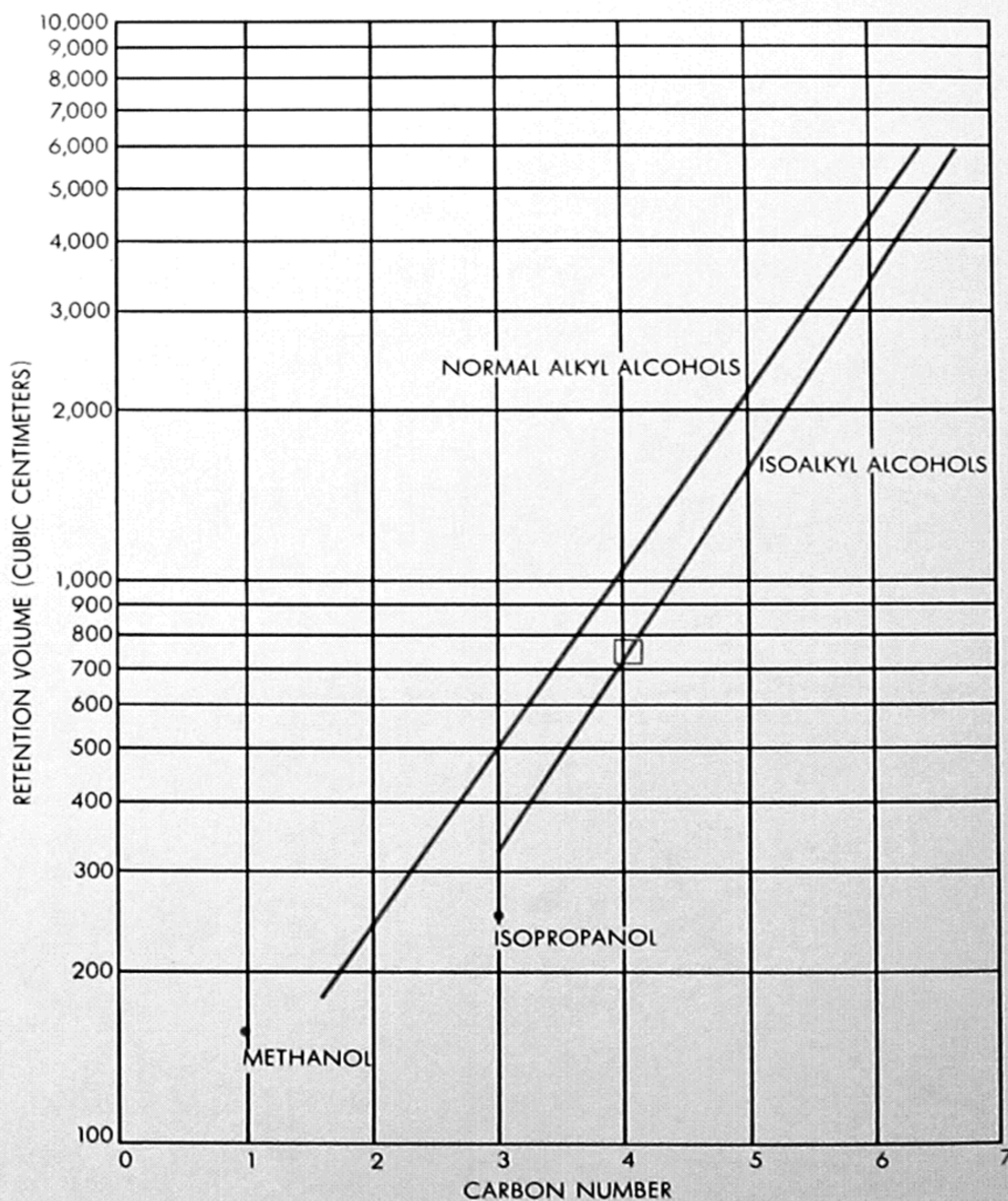
Before long one can reasonably expect that gas chromatography will be used not only as an analytical tool but also as a method of preparing ultrapure materials. Columns are now operating that can refine and separate the components in batches about one ounce in size. If the process can be made continuous, it should be able to provide laboratory chemicals and even commercial compounds having a wholly new order of purity.

Recently Donald E. Johnson, Sara Jo Scott and Alton Meister of Tufts University announced success in a long-sought goal: separation by gas chromatography of derivatives of the amino acids. The method may prove superior to the liquid-solid chromatographic method, employing a column packed with ion-exchange resins, that has been widely used for the analysis of the amino acids in proteins [see "The Chemical

Structure of Proteins," by William H. Stein and Stanford Moore; SCIENTIFIC AMERICAN Offprint 80].

A gas chromatograph of remarkable capabilities is now being developed by the Aerojet-General Corporation and the Jet Propulsion Laboratory of the California Institute of Technology; it will be placed aboard the *Surveyor* moon probe scheduled for launching in 1963. Upon reaching the moon the probe will pick up samples of lunar crust, grind them and feed them into a heater, where they will be pyrolyzed. The gas chromatograph will include two types of detector to analyze samples pyrolyzed at 150, 325, 500 and 1,000 degrees C. The heat-

er and chromatograph must consume no more than 10 watts of current. The whole apparatus must weigh no more than 11.5 pounds and must fit inside a box measuring 8 by 8 by 10 inches. Finally, to withstand the shock of landing on the moon, it must tolerate a deceleration force equivalent to 100 times the force of gravity on the earth. An important objective of the lunar chromatograph is to learn whether or not the moon's crust contains complex organic compounds of the type associated with living matter. This is an impressive assignment for an analytical tool that has come into general use only within the past five or six years.



COMPOUND IDENTIFICATION is often possible with gas chromatography when the sample is known to contain one of a limited number of compounds. For example, normal alkyl alcohols, whose carbon atoms form a straight chain, require a greater volume of carrier gas to drive them through a given column than isoalkyl alcohols having the same number of carbon atoms in a branched chain. Thus if an unknown four-carbon alcohol has a retention volume of 750 cubic centimeters of carrier gas (*square*), one can be sure it is isobutanol and not normal butanol, which would have a retention volume of about 1,000 c.c.

The Author

ROY A. KELLER is assistant professor of chemistry at the University of Arizona. Born in Davenport, Iowa, in 1928, he received a B.Sc. from Arizona in 1950 and an M.S. in chemistry in 1952. Keller then went to the University of Utah, where he studied under the direction of Henry Eyring. It was Eyring who first suggested that Keller do research in chromatography. Keller obtained his Ph.D. from Utah in 1956 and joined the faculty of Arizona later that year.

Bibliography

CHROMATOGRAPHY. Roy A. Keller, George H. Stewart and J. Calvin Giddings in *Annual Review of Physical Chemistry*, Vol. 11, pages 347-368; 1960.

PRINCIPLES AND PRACTICE OF GAS CHROMATOGRAPHY. Edited by Robert L. Pecsok. John Wiley & Sons, Inc., 1959.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

ULTRAHIGH VACUUM

by H. A. Steinherz and P. A. Redhead

In 1950 the limit of man-made vacuum appeared to be a pressure equivalent to 10^{-8} millimeter of mercury. A breakthrough in technique now provides pressures four orders of magnitude lower.

The prefixes "super-" and "ultra-" are rather freely used in scientific literature these days. These superlatives may prove embarrassing when it comes time to describe the next round of advances, but they do reflect an explosive rate of progress that has carried many techniques beyond what seemed to be the attainable limits only a short time ago.

The history of ultrahigh vacuum, a good case in point, goes back to 1950. Before that year high-vacuum practitioners had been improving their pumps, valves, seals and other components and had been reaching lower and lower pressures. In the middle 1940's they seemed to have reached a dead end. Using the best pumps and the most exquisite care in the design and operation of the systems, they could reach pressures approaching a hundred-millionth of a torr. (The torr, named for the 17th-century vacuum pioneer Evangelista Torricelli and now the standard unit in vacuum technology, is defined as the pressure necessary to support a column of mercury one millimeter high.) Apparently further refinements availed nothing. The gauges still indicated 10^{-8} torr. It was generally supposed that the pumps must somehow fail at this pressure. (The extremely small pressures with which this article deals are most conveniently expressed in negative powers of 10. The fraction $1/10$ is written 10^{-1} , $1/100$ as 10^{-2} and so on. Thus 10^{-8} is one hundred-millionth of a torr.)

It is worth pausing a moment to appreciate the achievement that a pressure of 10^{-8} torr represents. It is about one hundred-billionth the pressure of the atmosphere at sea level, which means that only one of every hundred billion air molecules originally present in the vacuum chamber is left after pumping.

At a pressure of one atmosphere each molecule travels, on the average, a few millionths of a centimeter before bumping into another molecule. At 10^{-8} torr, if it were not for the walls of the vacuum chamber, a typical molecule would travel almost 500,000 centimeters—some three miles—before encountering another. The pressure exerted by air at one atmosphere on a container (or on the mercury column of a pressure gauge) is the result of 3×10^{23} molecular impacts against each square centimeter of the container walls each second. At 10^{-8} torr the number of impacts is reduced, also by a factor of one hundred billion, to 3.8×10^{12} per square centimeter per second. This is still a lot of impacts, but not enough to hold up a column of mercury even one atom high. Obviously the definition of the torr given above no longer has any operational significance. The torr is nonetheless used to define these extremely low pressures, with the tacit understanding that it can be redefined in terms of a meaningful property such as the rate of impacts against a container wall, or the number of molecules per cubic centimeter (molecular density).

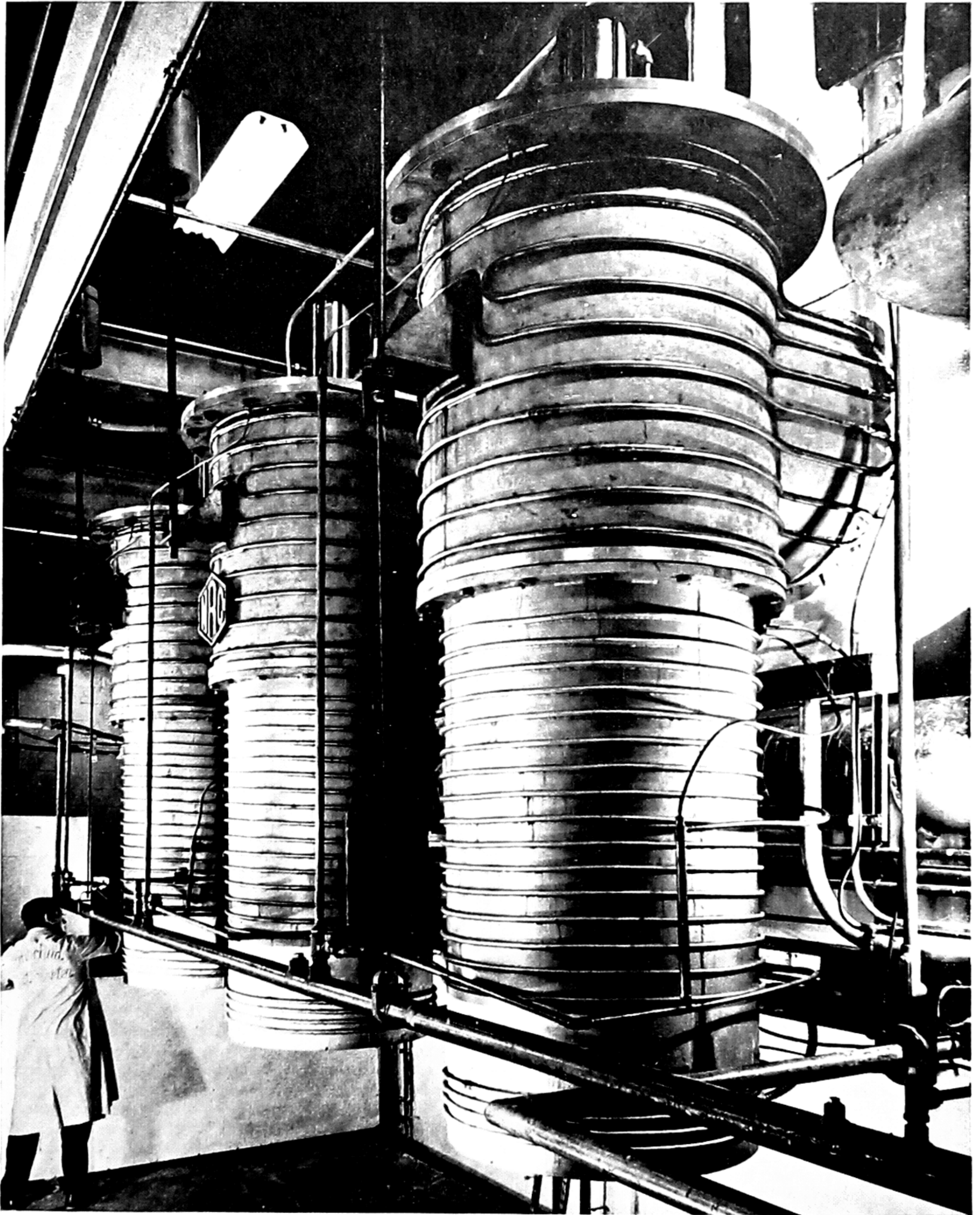
In 1947 Wayne B. Nottingham of the Massachusetts Institute of Technology suggested that the barrier at 10^{-8} torr was an illusion, caused by a failure in measurement rather than in pumping. The only instrument generally available that is capable of measuring pressures below about 10^{-4} torr is the ion gauge. The standard device in use during the 1940's consisted of a hot-wire cathode surrounded by a positively charged grid, which is in turn enclosed in an ion-collecting shell. The whole arrangement is enclosed in an envelope that is connected to the vacuum chamber, so that the gauge is in effect a part of the chamber. Electrons emitted from the central

cathode move rapidly toward the grid. In the course of their journey some of them collide with molecules of the gas to be measured, knocking electrons out of its molecules and producing positive ions. All the electrons are eventually collected on the grid. The positive ions move to the negatively charged collector, each one causing a tiny pulse of current to flow in the collector circuit. The number of ions produced depends on the density of the gas (that is, on the number of molecules per cubic centimeter), and so the collector current is an index to the molecular density.

Analyzing the operation of the ion gauge, Nottingham realized that there must be an additional process at work in it. Electrons bombarding the grid produce low-energy X rays. Some of this radiation would strike the ion collector and release electrons from its surface by means of the familiar photoelectric effect. So far as the current meter in the external circuit is concerned, the departure of a negative electron from the collector has exactly the same effect as the arrival of a positive ion. In short, the meter would register some current even if there were no gas whatever in the gauge and therefore no ions. Nottingham's calculations showed that this irreducible photoelectric current corresponded to a pressure of about 10^{-8} torr.

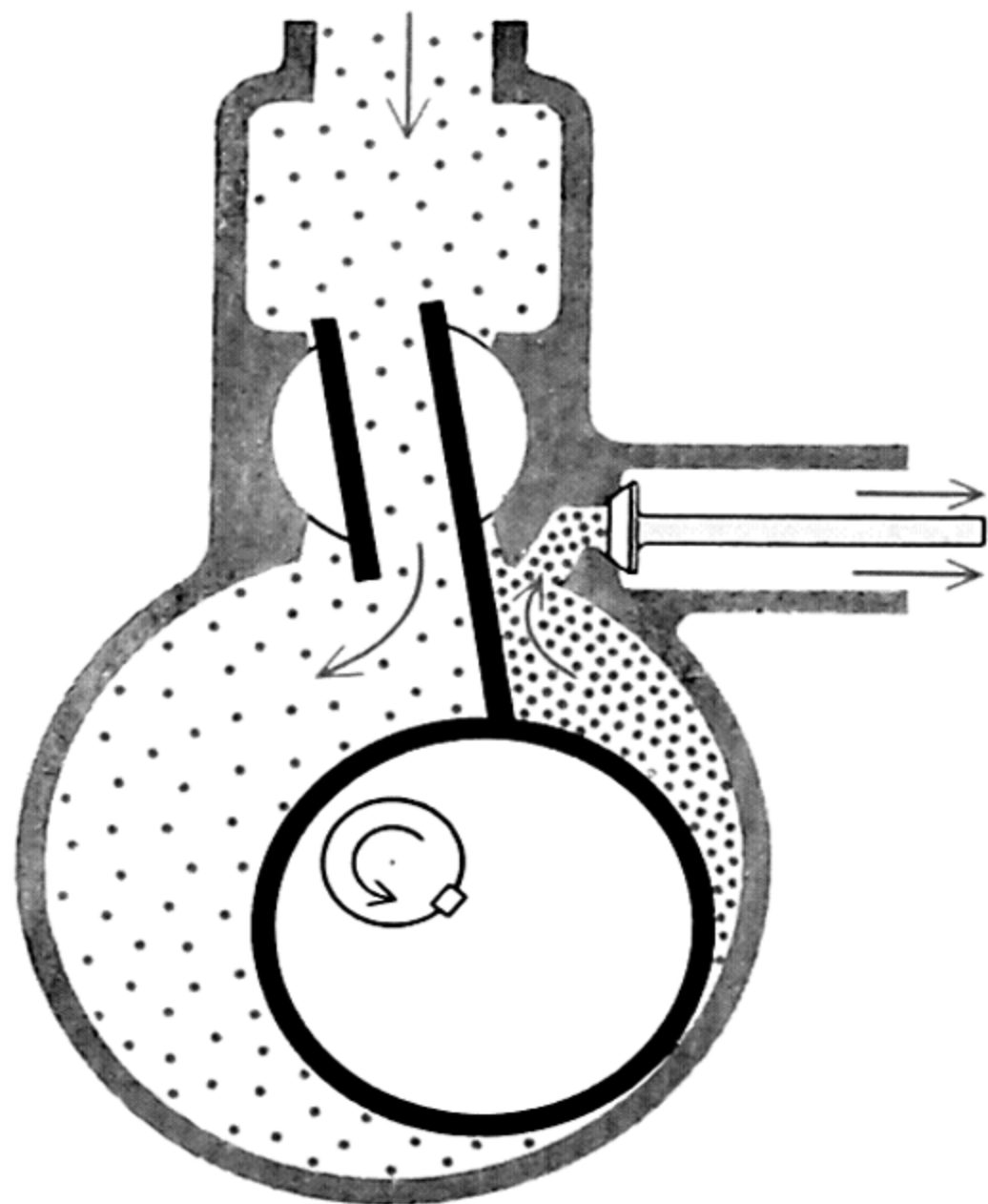
The Bayard-Alpert Gauge

A couple of years later Robert T. Bayard and Daniel Alpert, then at the Westinghouse Research Laboratories, hit on a simple modification of the ion gauge that both proved the correctness of Nottingham's analysis and greatly extended the limits of operation of the instrument. Essentially they switched the positions of the cathode and the ion

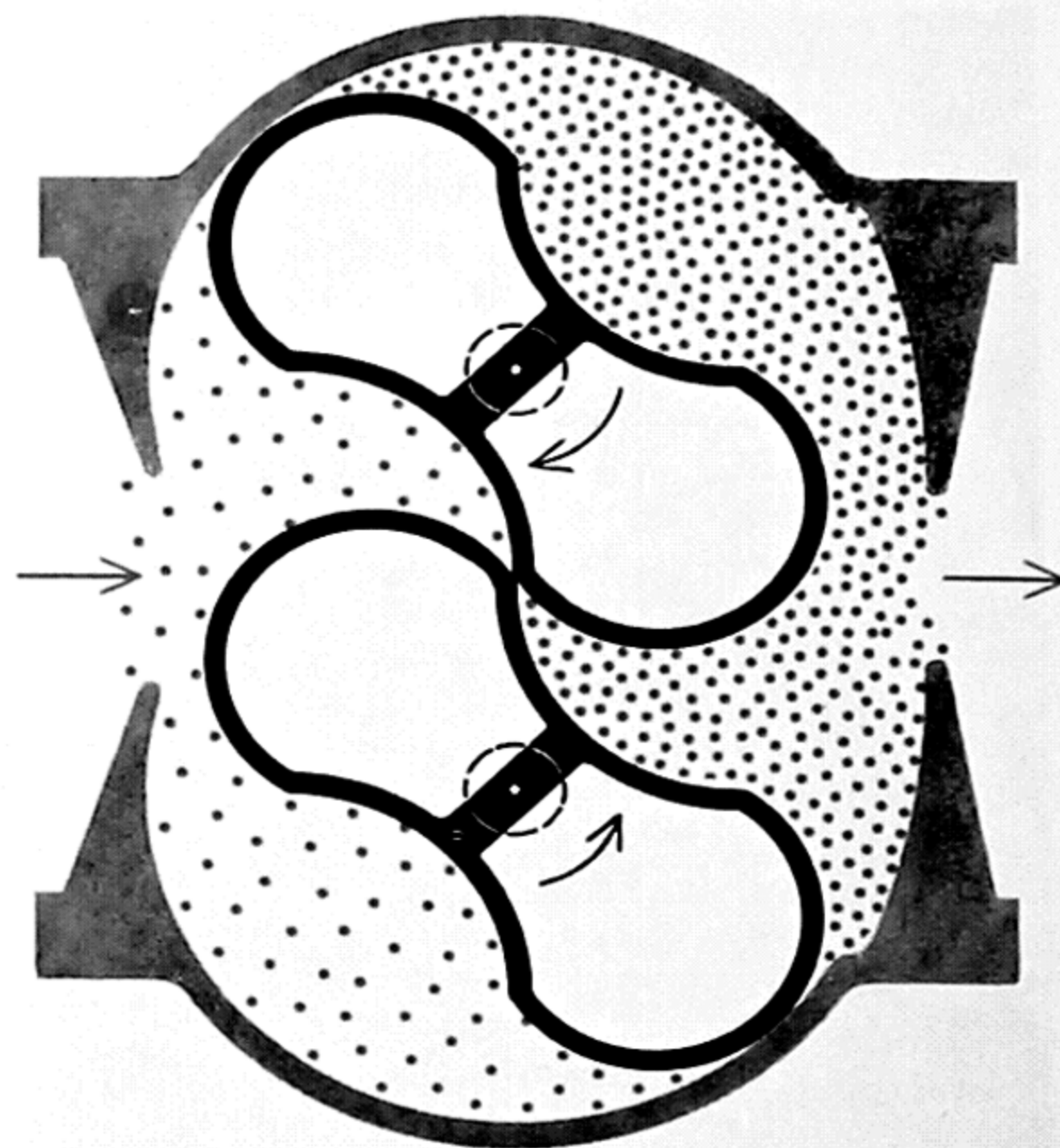


LARGE OIL-DIFFUSION PUMPS are located in Space Environments Laboratory of Fairchild Camera and Instrument Corporation's Defense Products Division at Syosset, N.Y. The three pumps,

each 32 inches in diameter, are teamed with three others to attain 10^{-8} torr (equivalent to a 300-mile altitude) in a 2,825-cubic-foot chamber in which space vehicles and sensors are to be tested.



MECHANICAL PUMPS are used in the first stage of evacuation. The rotary type (*left*) depends on the action of an eccentric rotating



piston to sweep gas out of a chamber. Blowers (*right*) contain two tightly fitted counterrotating lobes that trap and expel molecules.

collector. In the Bayard-Alpert gauge the cathode consists of a heated wire outside the grid, and the collector is a thin wire running down the axis of the instrument [see illustration at top left on page 660]. Carrying a negative charge, the collector picks up most of the ions formed in the gas. But because of its greatly reduced surface area it intercepts far less X radiation than a cylindrical plate does and therefore produces a much smaller photoelectric current. Bayard and Alpert showed that the residual current is equivalent to a pressure of about 10^{-11} torr.

With the new gauge Alpert was able in 1950 to break into the range of ultrahigh vacuum, below 10^{-8} torr. In fact, the instrument played a doubly crucial role. One of the difficulties with ion gauges had always been that they acted to change the very property they were supposed to measure. The reason is that the gas ions driven to the collector are trapped; hence they are removed from the vacuum chamber and the pressure drops. In other words, the ion gauge is also a pump.

Alpert proceeded to capitalize on the drawback. He pumped down a small glass chamber to almost 10^{-8} torr by

conventional means, and sealed off the chamber with a newly designed all-metal valve that required no organic sealing compound. (At these pressures conventional sealing compounds give rise to large amounts of gas.) Then he simply let the ion gauge continue to operate. Soon it was registering a pressure of 5×10^{-10} torr. Beyond this point the pressure would not go, although in principle the Bayard-Alpert gauge should have been capable of producing and measuring a pressure 10 times lower. Alpert then built a special type of mass spectrometer (which will be discussed later) to measure the pressures due to the various gases in the vacuum chamber. He found that the pressure limit in his system was set by the diffusion of helium atoms in the air through the glass walls of the chamber.

This work is as clear-cut an example of a scientific first as is ever likely to turn up. Yet, typically, once Alpert showed how to measure pressures in the ultrahigh-vacuum region it became clear that others had entered it before him. The new gauge demonstrated what some investigators had long suspected, that the diffusion pumps used in producing high vacuums could themselves pene-

trate the "barrier" at 10^{-8} torr. A review of the results of earlier work indicated that pressures not far from 10^{-9} torr had probably been attained as early as 1931.

Nevertheless, the opening of the age of ultrahigh vacuum must be dated from Alpert's remarkable experiments. After they had been published a growing number of workers entered the field, contributing further improvements to equipment and technique. Today a pressure of 10^{-12} torr is attainable, and still lower pressures are clearly in prospect. At 10^{-12} torr the molecular density is down to 33,000 molecules per cubic centimeter and the mean free path of a nitrogen molecule is about 50,000 kilometers, or 30,000 miles, long.

The Uses of Ultrahigh Vacuum

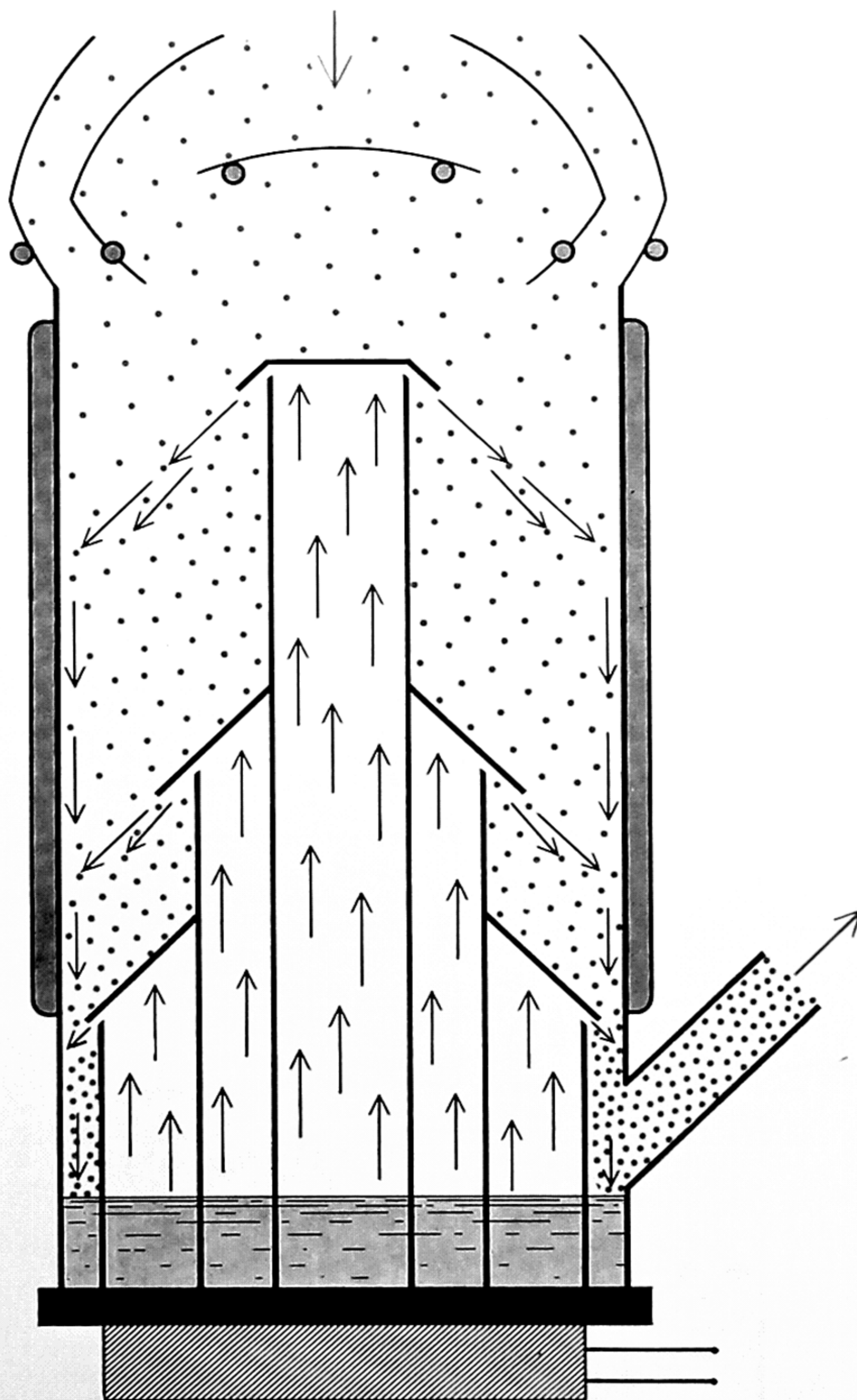
Before discussing the tools and methods of ultrahigh-vacuum technology, it seems appropriate to ask why they were developed. It is likely that sheer curiosity and the urge to explore would have produced them eventually. But there were far more practical lures, in both fundamental and applied research, to attract men and money to the task.

The original impetus toward ultrahigh

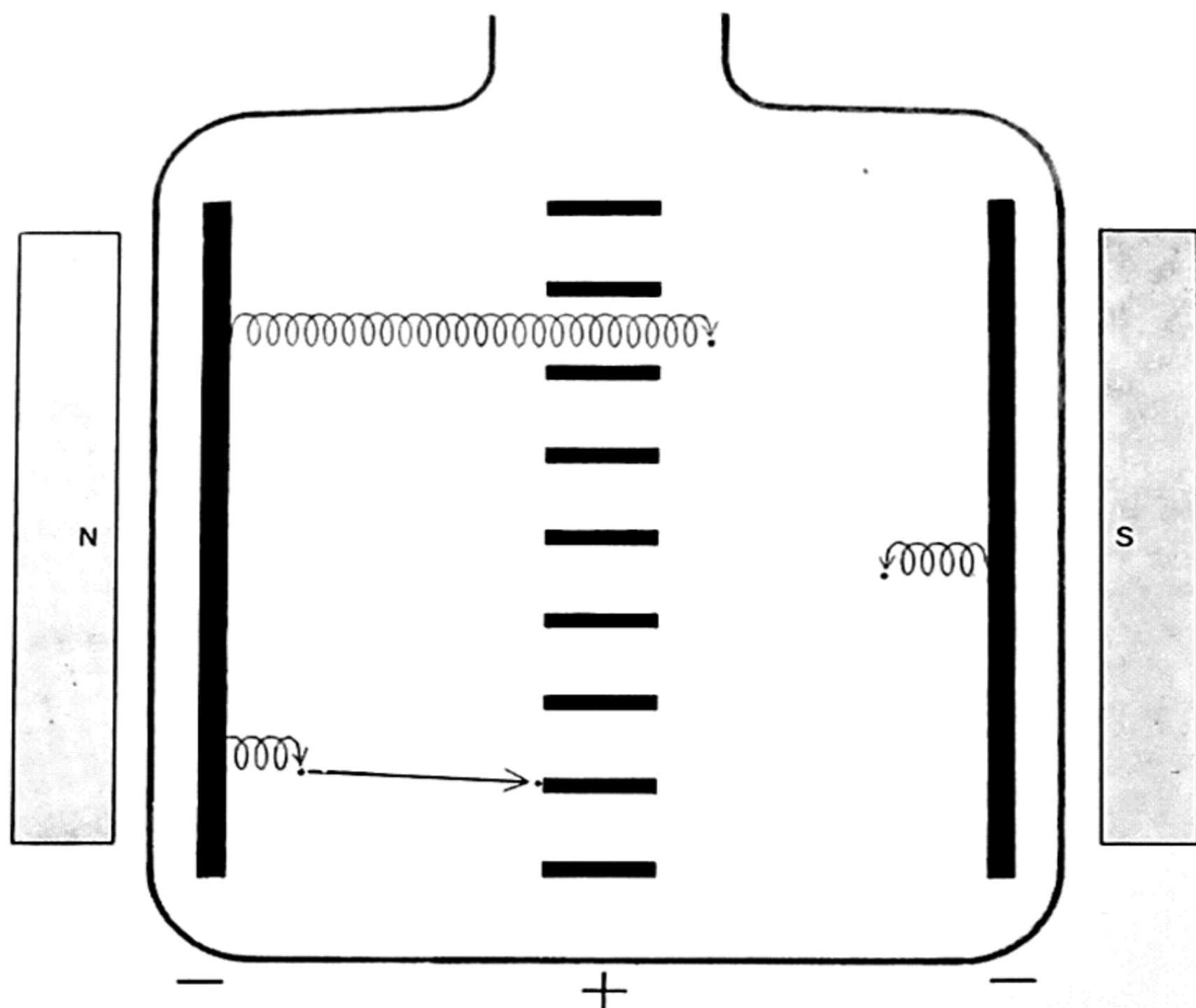
vacuum came from the requirements of experimenters studying the physics and chemistry of solid surfaces. Most of the properties of such surfaces can be studied only on surfaces that have been cleaned of adsorbed gases and will stay clean long enough for the appropriate measurements to be made. Metals and semiconductors can be cleaned on an atomic scale by heating them in a vacuum or by spraying their surfaces with a low-energy discharge of an inert gas. How long the pristine surface lasts then depends on the rate at which it is bombarded by adsorbable molecules from the surrounding gas. At a pressure of 10^{-6} torr, considered a good vacuum only a few years ago, a layer of gas one molecule thick will form on a metal surface in about one second. If the surface has been cleaned by heating in a vacuum of 10^{-6} torr, it is usually completely covered before the sample has cooled to its initial temperature, and before many of the desired measurements can be carried out. At 10^{-9} torr the "monolayer time" increases to about 20 minutes, and it becomes possible to observe the clean surface for a reasonable length of time.

A remarkably wide variety of the properties of a material are affected by gases adsorbed on its surface. One obvious example is the force of friction. Indeed, the true friction between metallic surfaces has so far been measured for only a few metals. The standard handbook figures really refer to surfaces lubricated by adsorbed gas layers. Again, the emission of electrons from a solid surface (by thermionic emission, photoelectric emission or other processes) is sensitive to surface contamination, and ultrahigh-vacuum techniques are now being widely used to study the phenomenon. Some electrical properties of semiconductors are also strongly affected by adsorbed gas. Finally, the process of adsorption itself is more amenable to study when it can be observed in slow motion, so to speak, in an ultrahigh-vacuum chamber.

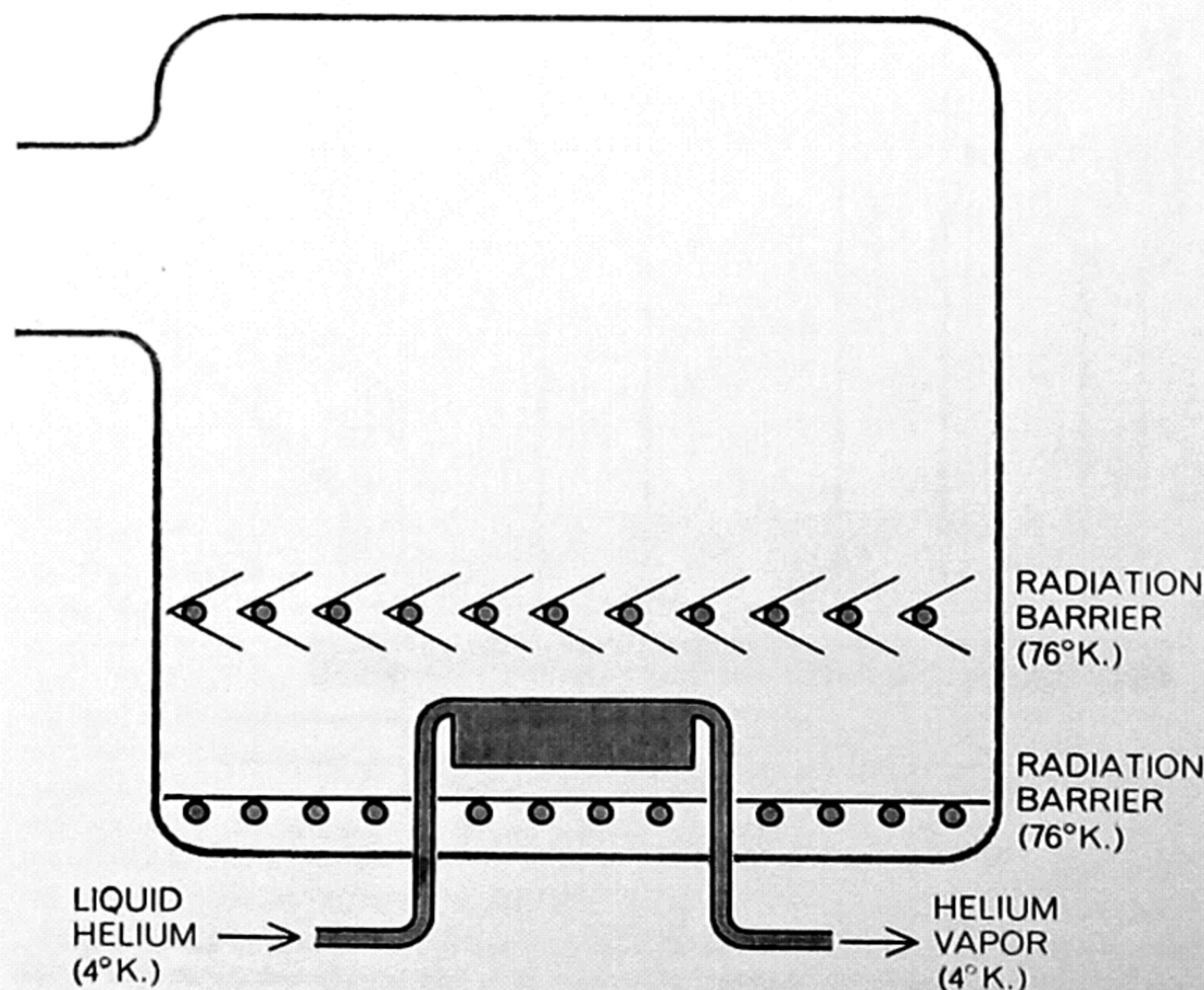
In any experiment involving gases of high purity at low pressures, ultrahigh-vacuum techniques are indispensable. For example, if it is required to maintain a purity of one part per million in a gas sample at a pressure of 10^{-3} torr, the vacuum chamber must be evacuated to less than 10^{-9} torr before the sample is introduced. If the experiment involves an electrical discharge, the difficulties are compounded. When radiation or gas ions from the discharge strike the vacuum chamber, they release contam-



DIFFUSION PUMP is the work horse of high vacuum. A liquid, usually oil or mercury, is vaporized by a heater at the bottom. The vapor rises and is deflected downward (*black arrows*) as a high-speed jet that entrains gas molecules (*colored dots*) diffusing out of the vessel above and carries them out of the pump. Back-diffusion of pumping vapor into the chamber is inhibited by baffles, cooled by coils (*gray circles*), on which the vapor freezes.



SPUTTER ION PUMP works by ionizing gas molecules and removing them from the chamber (top) to be evacuated. Electrons emitted by the cathode plates are accelerated (in spirals because of the magnetic field) by the anode. When they strike gas molecules, they create positive ions (open colored circles). These are attracted to the negative collector plates, from which fresh titanium metal is "sputtered" to provide clean pumping surfaces.



CRYOGENIC PUMP is another device that immobilizes gas molecules and thus removes them from a system. It does so by condensing the molecules on a surface cooled by liquid helium. The V-shaped baffles prevent heat radiation downward but allow molecules to pass.

inating gases from the walls.

Maintaining clean surfaces is essential not only in laboratory studies but also in manufacturing many of the new miniaturized solid-state devices. These are made by depositing extremely thin layers of different materials, one over another on a base plate. The composition and dimensions of the layers must be precisely controlled, which is possible only under ultrahigh-vacuum conditions.

Conventional vacuum-tube electronics furnished one of the major incentives for the development of high-vacuum technology. With a few exceptions normal vacuum techniques are still adequate for tubes. In any vacuum tube, however, residual gases produce a number of undesirable effects, including increased noise, increased grid current and damage to the cathode through bombardment by positive ions. When the standards of performance are exacting, as in low-noise or high-power vacuum tubes, the pressure in the tube envelope must be reduced to the ultrahigh-vacuum range. High-power klystron tubes, for example, require a vacuum of better than 10^{-9} torr.

A new class of electron tube, still under development, obtains its current through the "field emission" of electrons due to high electric fields at a fine, unheated metal point. Tube failure through filament breakage is thereby avoided. To obtain stable emission requires extremely low pressures. W. P. Dyke at the Linfield Research Institute in Oregon has built tubes that maintain stable emission for thousands of hours but only at a pressure of 10^{-12} torr. To maintain this vacuum the tube envelope must be made of a special aluminosilicate glass that has a low permeability to the helium in the atmosphere.

In all the cases mentioned so far the amount of space that must be evacuated is not very large—in the cubic centimeter range. There are two major applications in which the volume requirements are considerably greater.

One of them is in the machines with which physicists seek to solve the problem of how to control the release of power from the thermonuclear reactions of the light isotopes deuterium and tritium. An important obstacle to the achievement of thermonuclear power is the loss of energy by radiation from the hot gas, or plasma. Any contamination of the gas by the atoms of heavier elements increases the loss enormously, since these atoms radiate much more strongly than lighter ones do. When one of the early

experimental machines was operated after evacuating it to 10^{-6} torr, the lowest pressure then attainable, the radiation loss was so great that the plasma temperature reached only a tenth of the calculated value. To reduce the loss to acceptable levels requires pressures in the ultrahigh-vacuum range. Recent improvements in technique have made it possible to maintain pressures in the range of 10^{-10} torr, where contamination is no longer a major problem. If practical thermonuclear power is ever achieved, the reactors will have to be very large. From the viewpoint of the vacuum engineer even today's experimental devices are big enough. For example, the evacuated volume in the Model C "stellarator" at Princeton University is about half a cubic meter. Not only must this chamber be pumped down to 10^{-10} torr before the plasma is introduced but also the pressure of the contaminants must be maintained at this value as impurities are released from the walls of the chamber by the action of the high-energy discharge.

The Simulation of Space

The most insistent, as well as the most stringent, demands on ultrahigh-vacuum technology come from the space-exploration program. At 450 miles above the surface of the earth the pressure of the atmosphere is only 10^{-9} torr, and at 1,200 miles it falls to about 10^{-11} torr, approaching the limit now generally attainable in the laboratory. In interplanetary space the pressure is estimated to be on the order of 10^{-16} torr. This corresponds to a density of about four molecules per cubic centimeter.

Under these ultrahigh (and ultraultrahigh) vacuum conditions materials may behave in unexpected ways, and physical and chemical processes may be radically altered. For example, graphite, an excellent lubricant at atmospheric pressure, becomes an abrasive below 10^{-6} torr. Heat transfer, fluid flow, dielectric behavior and other phenomena change drastically as pressure is decreased.

Not all the changes are necessarily harmful. The resistance of metals to certain types of fatigue increases greatly at low pressure. A strip that will break after a few flexings under ordinary conditions can be bent back and forth for hours in an ultrahigh-vacuum chamber. It appears that in an ultrahigh vacuum tiny cracks that form at each bending reweld themselves when the strip is bent the other way. In air at atmospheric

pressure the cracks are covered with an oxide coating as soon as they open. The metal cannot join together again, and as a result the cracks grow larger each time the piece is bent.

To test equipment designed for space vehicles, simulation chambers have been built at a number of laboratories. Small or moderate-sized enclosures that can be evacuated to the lowest pressures now attainable suffice for examining fundamental processes and the properties of materials. But to check the performance of complete components requires chambers of enormous size. The biggest ones go up to some 1,000 cubic meters. Even these huge spaces are now maintained at pressures near 10^{-8} torr.

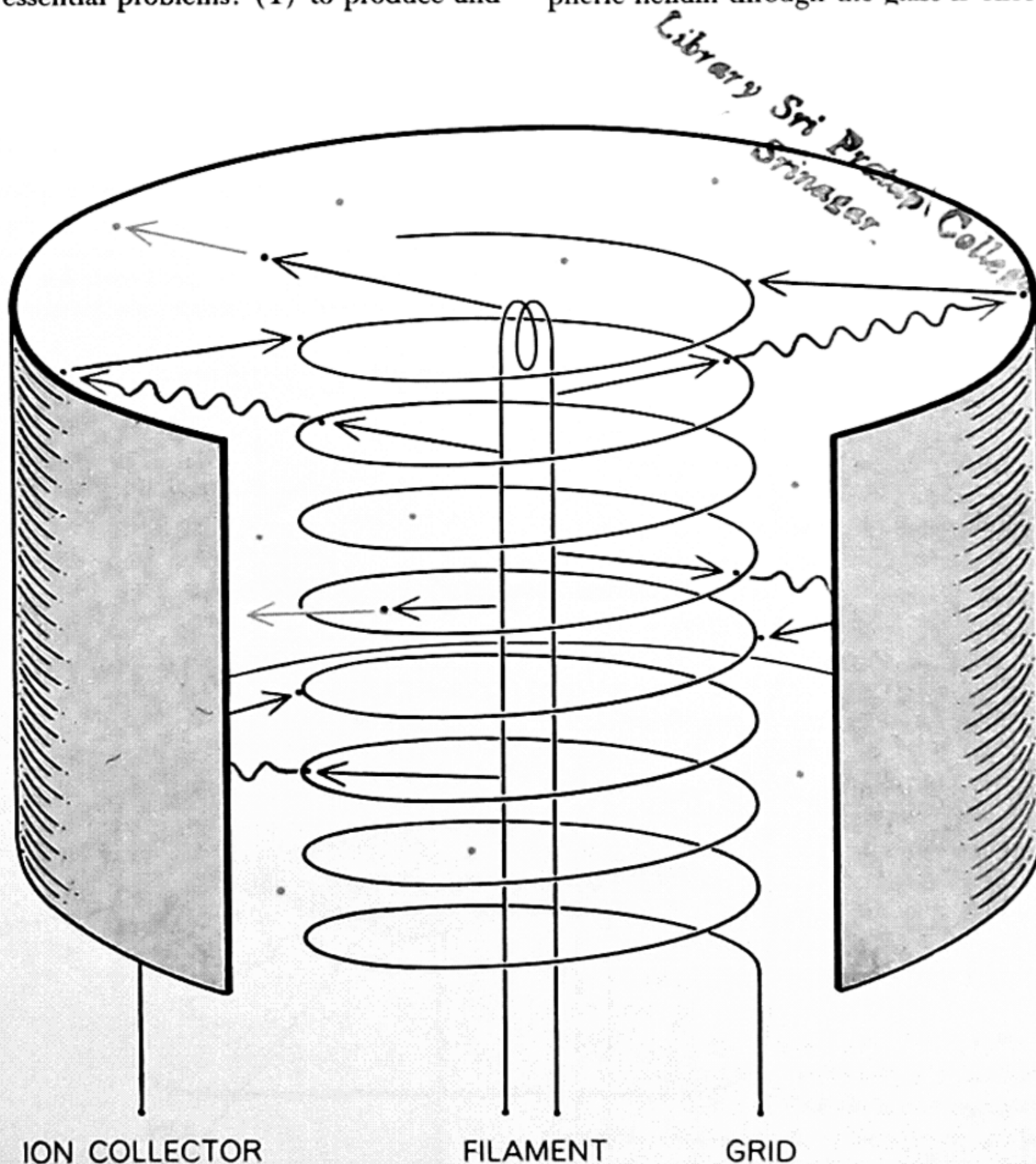
Producing the Ultrahigh Vacuum

Ultrahigh-vacuum workers face two essential problems: (1) to produce and

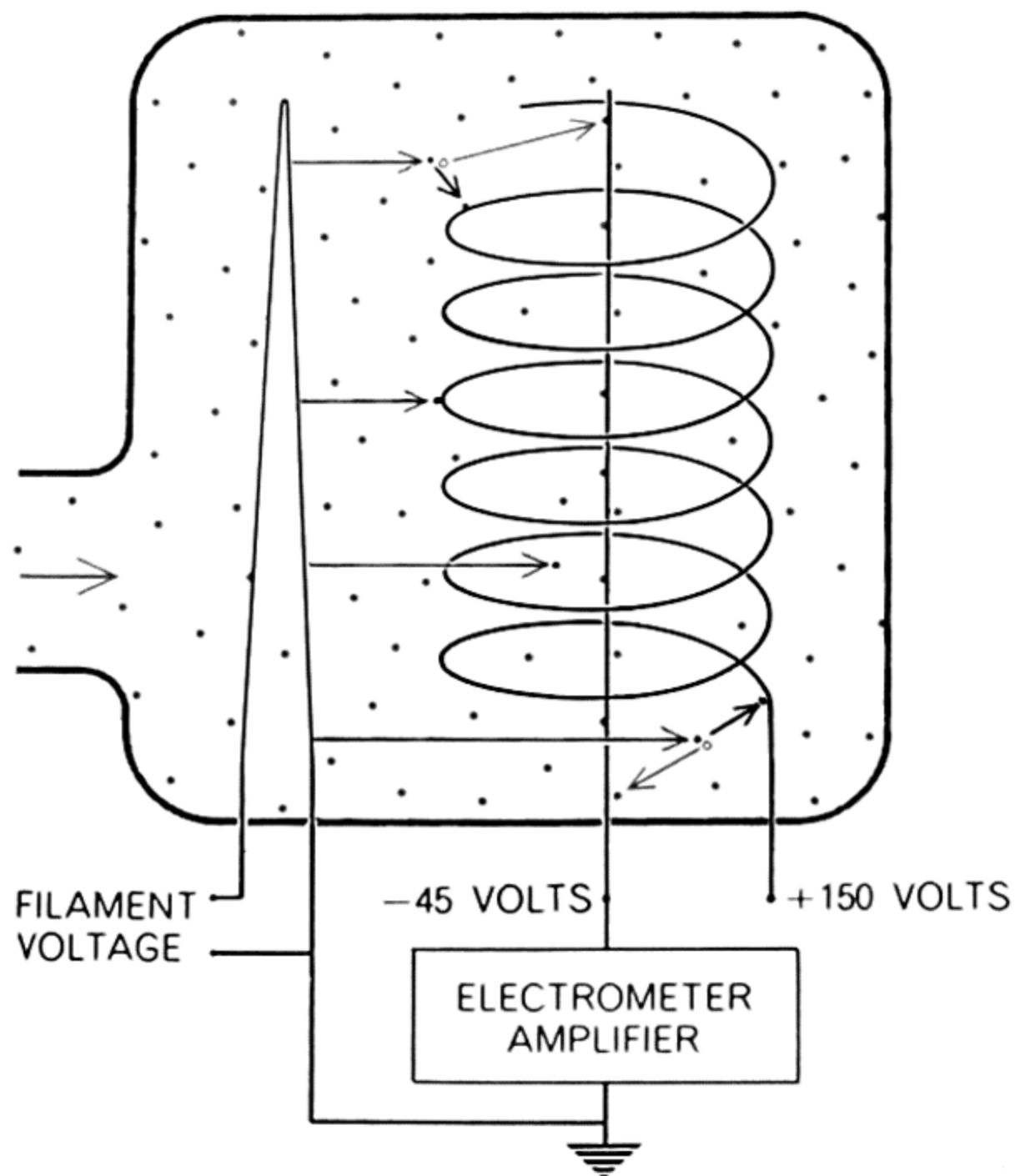
maintain pressures below 10^{-8} torr and (2) to measure the pressure they have achieved. So far as the first problem is concerned the solution in every case represents a balance between the capacity of the pump and the system gas load: the rate at which gas continuously enters the vacuum chamber from the various parts of the system (including the pump).

Each part contributes its share, which the vacuum engineer seeks to minimize. Following Alpert's lead, the valves in ultrahigh-vacuum systems are made of metal and only low-vapor-pressure sealing compounds are used.

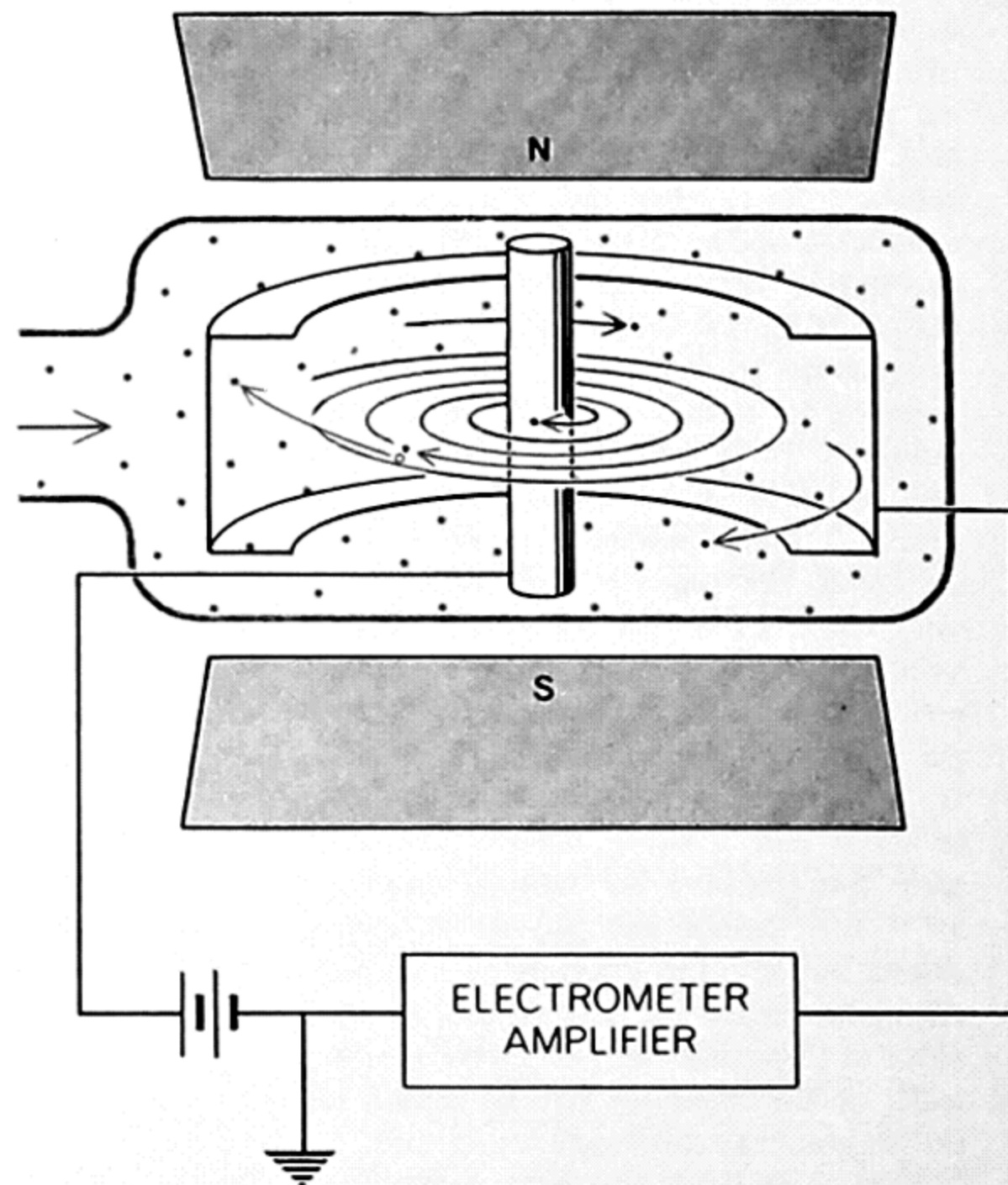
In small systems where the experimenter wants to observe what is going on inside or to be able to change his set-up easily, the ultrahigh-vacuum chamber and its connecting tubing are still made of glass. The penetration of atmospheric helium through the glass is offset



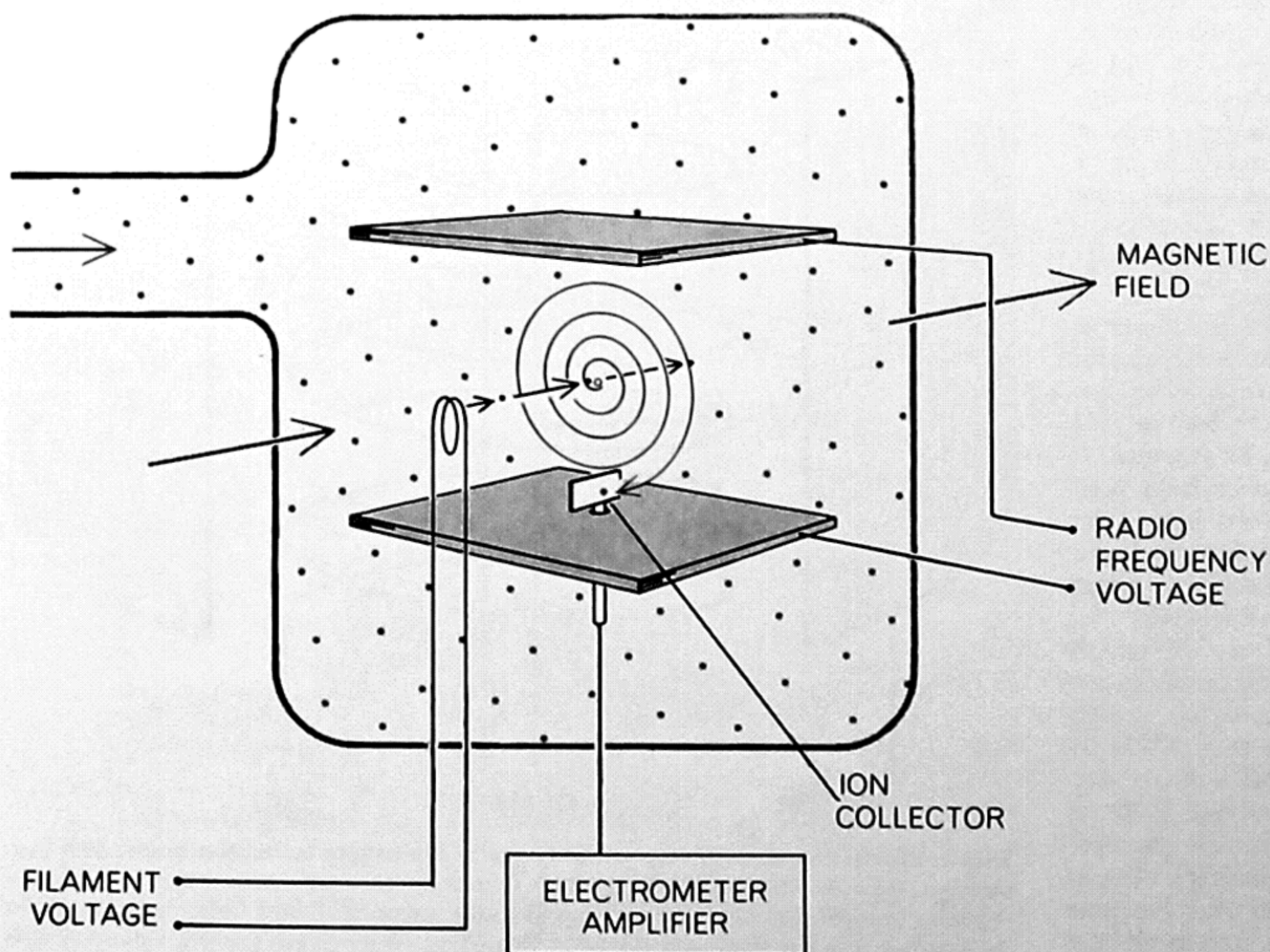
X-RAY PROBLEM made it impossible for early ion gauges to register below 10^{-8} torr. Electrons from the filament created positive ions (colored circles) that struck the collector and were counted. But electrons reaching the grid produced X rays (wavy arrows). When the X rays struck the large-area collector, they liberated electrons, causing a photoelectric current that could not be distinguished from the current resulting from ion impact.



BAYARD-ALPERT GAUGE avoided the X-ray problem by putting the heated-filament cathode outside the grid and making the collector a thin axial wire. The negatively charged collector still gathers positive ions, but because of its small area it intercepts fewer X rays and therefore emits a smaller photoelectric current.

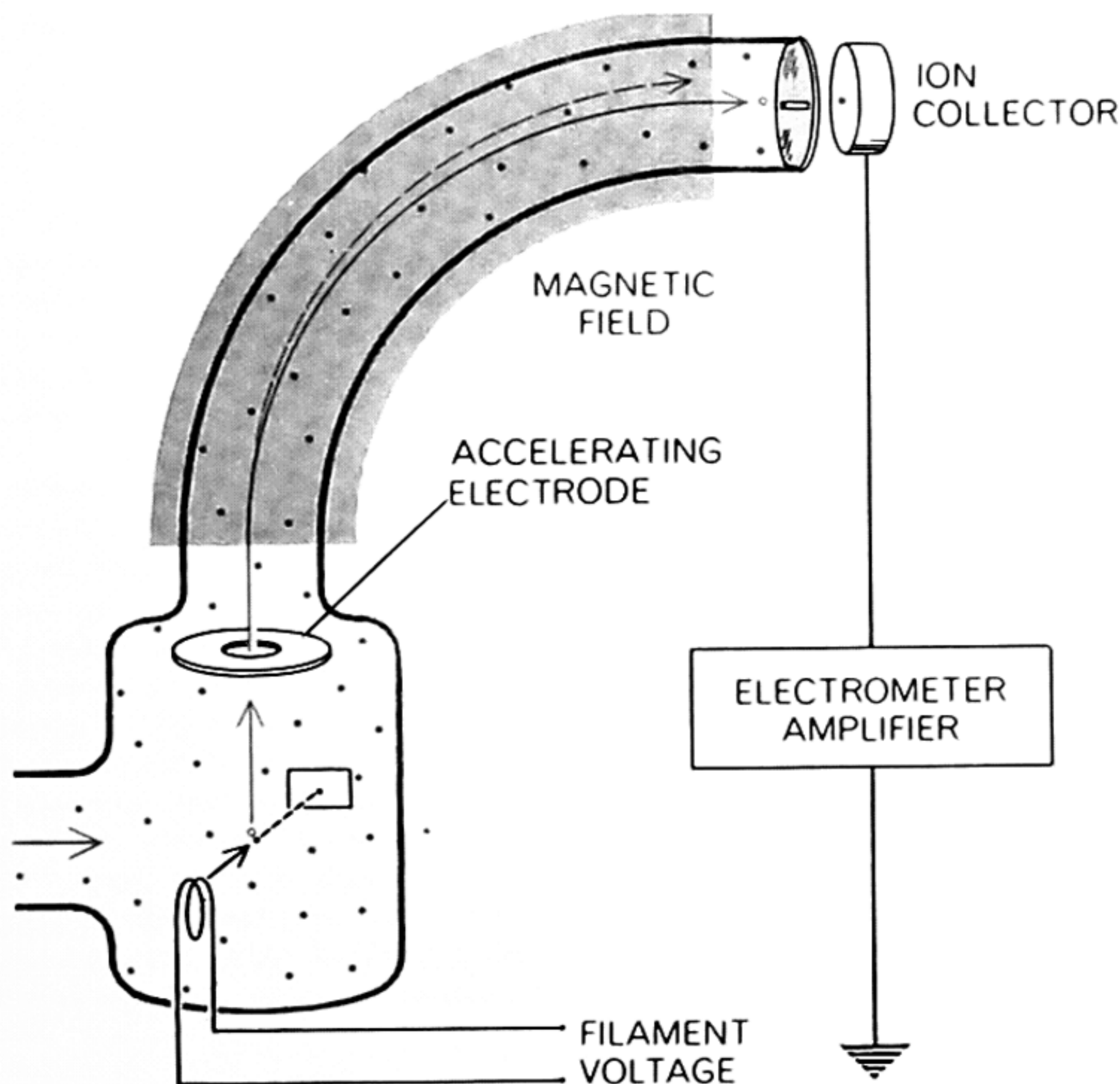


INVERTED MAGNETRON, a cold-cathode gauge, produces electrons by applying a high voltage to unheated electrodes. In such a gauge X rays are proportional to pressure, so no spurious current is produced. Electrons spiraling in toward the central electrode ionize gas molecules, which are collected on the curved cathode.

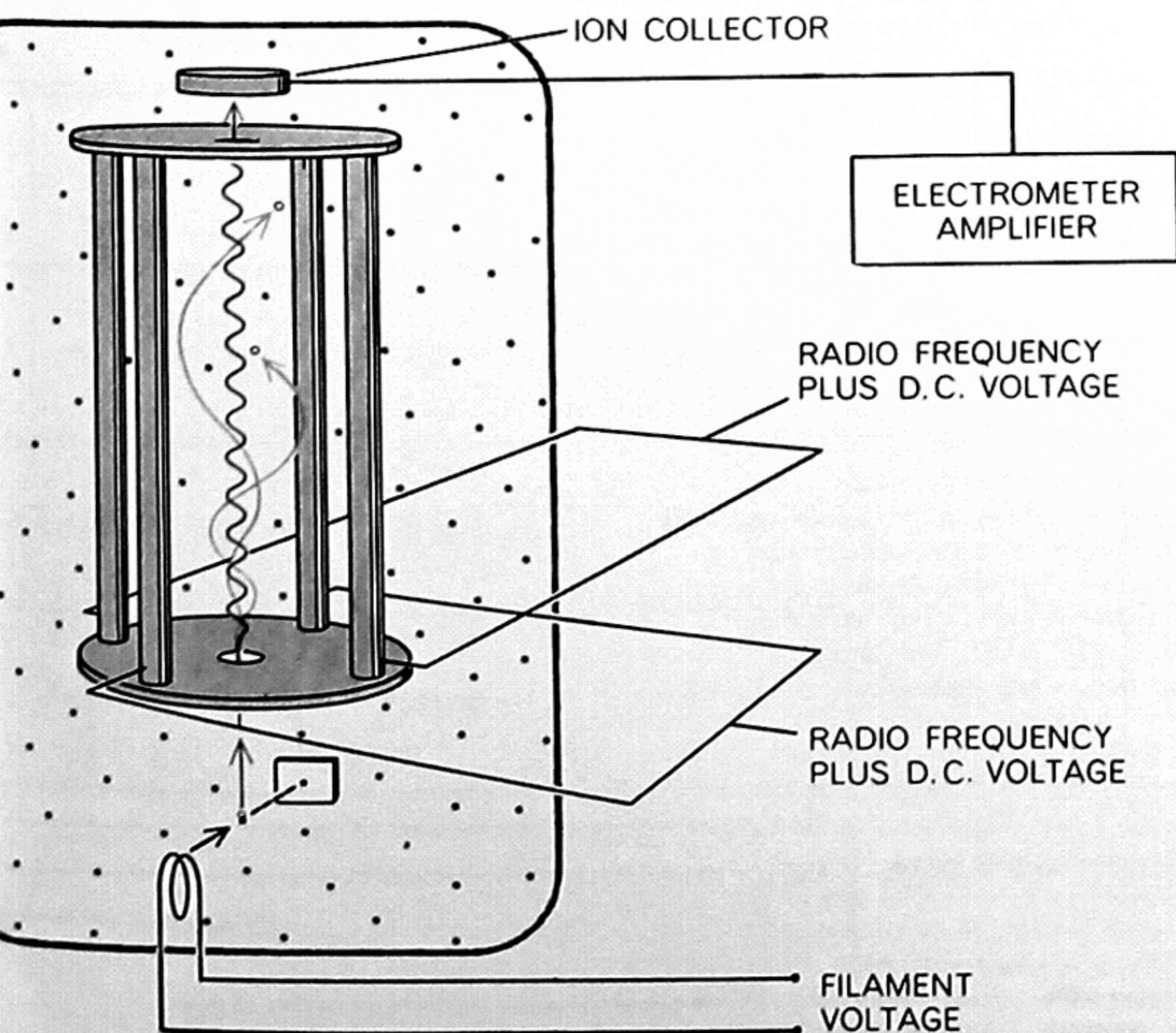


OTHER MASS SPECTROMETERS are diagramed. In the Omega-tron (*left*) a radio-frequency voltage is varied to whirl positive

ions of different masses into spirals that strike the collector. In the massen filter (*right*) the ion beam is accelerated along the axis of



MASS SPECTROMETER can measure the partial pressure of individual gases instead of just total pressure. In this gauge, a modification of a standard spectrometer, gas molecules are ionized by an electron gun (*bottom*) and then accelerated through a magnetic field. By varying the magnetic or electric field one can direct ions of different masses through the slit to strike the collector and so measure the density of one constituent gas at a time.



four rods carrying both direct-current and radio-frequency voltages. At each particular voltage ions of only one mass spiral up the axis to the collector; the others hit the rods.

by a pump of sufficiently high capacity. Moreover, as has been mentioned, the composition of the glass itself can be altered to reduce its permeability.

All large ultrahigh-vacuum systems, and nowadays many small ones, are built of metal, usually stainless steel. In the early days (10 years ago) designers had to weld or braze all joints to avoid excessive leakage. This equipment was hard to modify, and even introducing experimental samples into the chamber was awkward. With the recent development of improved metal gaskets and of certain elastomers such as Viton, which vaporizes very slowly even in an ultrahigh vacuum, bolted joints have become practical and metal vacuum systems are much more flexible.

Whatever the system is made of, all its parts must be able to withstand baking for hours at a temperature of several hundred degrees centigrade to drive off gas adsorbed on the surfaces and absorbed within the solid materials. Without this preparation the gas would steadily evolve into the chamber as pumping started and the pressure began to drop. To reach an ultrahigh vacuum would take an inordinately long time.

Even after the most careful preparation, desorption from the walls of the system constitutes a major source of the residual gas that keeps dribbling into an ultrahigh-vacuum system. An idea of the dimensions of the problem is provided by a typical small stainless steel system with walls one millimeter thick and a surface area of 100 square centimeters. With good technique, the rate of evolution can be reduced to about 10^{-10} torr-liter per second. (This is a rate that would raise the pressure of an absolutely empty one-liter chamber to 10^{-10} torr in one second.) Inflow through microscopic leaks in valves and joints would contribute the same amount of gas. Permeation of helium through the walls would add only about a hundredth as much; in metal systems this permeation is negligible. A small laboratory ultrahigh-vacuum pump, with a capacity of 10 liters (a little more than 10 quarts) per second, can maintain this chamber at 2×10^{-11} torr. (The rating of a pump must be referred to the pressure range in which it operates. Working at 2×10^{-11} torr, the pump in the above example can remove in one second the number of gas molecules that a 10-liter chamber contains at 2×10^{-11} torr.)

In principle the same figures should hold for larger systems so long as the

pumping rate increases proportionately with the wall surface. Space-simulation chambers often have areas of a million square centimeters or even more. Typically their pumping systems are rated at 100,000 liters per second. With the area multiplied 10,000 times and the pumping speed increased by the same factor, a pressure of 2×10^{-11} torr should still be attainable. In practice, however, unavoidable leaks at the many large gaskets and seals have limited the pressure in the big chambers to about 10^{-8} torr.

The Pumps

The evacuation of any chamber, regardless of size or eventual pressure, begins with mechanical pumping. The pump is almost always the familiar rotary type, in which gas is swept out by an eccentric rotating piston [see illustration at left on page 656]. Sometimes, particularly in large systems, the rotary pumps are joined in series with blowers,

which remove gas by the action of two counter-rotating lobes [see illustration at right on page 656]. Mechanical pumps bring the chamber down to about a thousandth of a torr. Then the ultrahigh-vacuum pumps take over.

Today diffusion pumps are the work horses of high vacuum, as they have been since their invention 40 years ago. Improvements have now extended their usefulness through the whole of the present ultrahigh-vacuum range.

The only moving part in a diffusion pump is a high-speed jet of oil vapor or some other vapor, directed away from the opening of the vacuum chamber [see illustration on page 657]. Gas diffusing out of the chamber is entrained in the jet and swept along with it out of the pump. The device is simple and fast and can be made as big or as small as is desired. Its major drawback—the factor that eventually limits the pressure it can reach—is the diffusion of vapor from the pumping fluid back into the vacuum chamber.

In the past few years the limit has been pushed back through a number of technical advances. Pump jets have been redesigned to allow higher boiler pressures and to reduce migration of the fluid into the vacuum chamber. New pump oils are available, consisting of polyphenyl ethers or liquid silicones, that have vapor pressures of less than 10^{-9} torr at room temperature, 10 to 100 times lower than those of the best oils available five years ago.

No matter how low the volatility of the fluid is, some of it is bound to back up in the pump. A good part of the contaminant, however, can be kept out of the vacuum chamber with baffles. The most common arrangement consists of a series of vanes, placed at the entrance to the chamber and cooled to the temperature of liquid nitrogen, on which the oil vapor freezes. Recently “molecular sieve” materials such as zeolite have demonstrated effective baffling action. With the help of zeolite filters operating at room temperature, diffusion-pumped

	ATMOSPHERIC PRESSURE	BEGINNING OF HIGH-VACUUM REGION	BEGINNING OF VERY-HIGH- VACUUM REGION	BEGINNING OF ULTRAHIGH- VACUUM REGION		
PRESSURE (TORR)	760	10^{-3}	10^{-6}	10^{-8}	10^{-9}	
NUMBER OF MOLECULES PER SECOND BOMBARDING EACH SQUARE CENTIMETER OF CONTAINER WALLS	3×10^{23}	4×10^{17}	4×10^{14}	4×10^{12}	4×10^{11}	
MEAN FREE PATH OF ONE MOLECULE BETWEEN COLLISIONS WITH OTHER MOLECULES (CENTIMETERS)	6.5×10^{-6}	5	5×10^2	5×10^5	5×10^6	
NUMBER OF MOLECULES PER CUBIC CENTIMETER	2×10^{19}	3×10^{13}	3×10^{10}	3×10^8	3×10^7	
TYPE OF PUMPS USED	MECHANICAL: ROTARY AND BLOWER		DIFFUSION ————— ION ————— CRYOGENIC —————			

CHARACTERISTICS OF VACUUMS are charted, beginning with atmospheric pressure and extending to the currently practical

laboratory limit. Values are given approximately, in orders of magnitude. Since “pressure” as such cannot be measured in extremely

systems have been held at 10^{-10} torr for more than 100 days. Refrigerating the zeolite extends its useful life even further and appears to eliminate oil contamination completely.

The decomposition of the pumping fluids has limited oil diffusion pumps to about 10^{-10} torr. Better results have been achieved by replacing oil with mercury. Small mercury diffusion pumps have now reached all the way down to 10^{-12} torr.

In spite of the many advantages of the diffusion pump, it is often replaced in small laboratory systems by the ion pump. Basically this device operates in the same way as the gauge that Alpert used to attain the first certified ultrahigh vacuum: ions, produced by electrons bombarding the gas molecules, are driven by an electric field to a collector, where they are adsorbed. Note that the ion pump works on a quite different principle from that of the diffusion pump. Instead of taking gas entirely out of the system, ion pumps immobilize it

within the system. Herein lies one of their weaknesses: low capacity due to the saturation of the collecting surface.

Alpert's gauge had a capacity of only a small fraction of a liter per second. By designing the device primarily for pumping, the speed can be increased to many liters per second [see top illustration on page 658]. Containing no foreign pumping fluid, the ion pump contributes far less contamination than the best of the diffusion pumps. It is therefore better suited to operate unattended for long periods.

When the action of an ion pump is analyzed, it turns out to depend on two different processes. Some of the ions reaching the collecting surface are held there by adsorption. Others, however, react chemically with the collector material, forming stable compounds. Chemically active gases are pumped by both mechanisms; inert gases, only by the former. In most ion pumps the collector is made of titanium. The steady hail of ions "sputters" the surface, continuously providing fresh layers of clean titanium metal with which the chemically active gases combine.

All ion pumps suffer to some extent from the same pair of defects. First, they pump chemically active gases much faster than they do inert gases. Second, they re-emit a small fraction of the gas previously pumped. The re-emitted gas acts as an additional source in the vacuum system. Moreover, if one gas is pumped initially and then a second gas is introduced, the pump will remove the second while re-emitting the first. New designs have considerably reduced the size of this effect.

The Cryogenic Pump

Another type of immobilization device now coming into prominence is the cryogenic pump. It removes gas simply by condensation on a very cold surface [see bottom illustration on page 658]. Cryogenic pumps can be used in any position, and their speed is not restricted by connecting tubing. These advantages make them particularly valuable for large space-simulation chambers.

When a surface is cooled to the boiling point of helium (4.2 degrees centigrade above absolute zero), all gases are adsorbed, including helium. Pressures of about 10^{-13} torr have already been obtained in small glass systems by partially immersing them in liquid helium. From a study of the adsorption process J. P. Hobson of the National Research

Council of Canada has estimated that a half-liter glass sphere evacuated to a pressure of 5×10^{-10} torr by normal methods and then completely immersed in liquid helium should reach about 10^{-30} torr. For better or worse there is no pressure gauge capable of checking this pressure. The calculations indicate, however, that it should be possible, at least in small systems, to produce pressures in the laboratory as low as those thought to exist in interplanetary space.

At the moment, then, pumping has outstripped measurement. But the gauge too has been improved a great deal beyond where Bayard and Alpert left it. Their hot-wire design becomes more effective when the gauge is placed between the poles of a strong magnet. Instead of moving in straight lines the electrons are forced into spirals, which increases their length of travel and therefore their chance of ionizing gas molecules. Changes in the shape and placement of electrodes have further reduced the X-ray current. Nevertheless, the best heated-filament gauges now generally available are only slightly better than the Bayard-Alpert type: their lower limit is about 10^{-10} torr, as opposed to 5×10^{-10} . Recently James M. Lafferty of the General Electric Research Laboratory has built a hot-wire gauge with a magnetic field, capable of measuring pressures as low as 4×10^{-14} torr. Attached to an electron multiplier, its range has been extended even further.

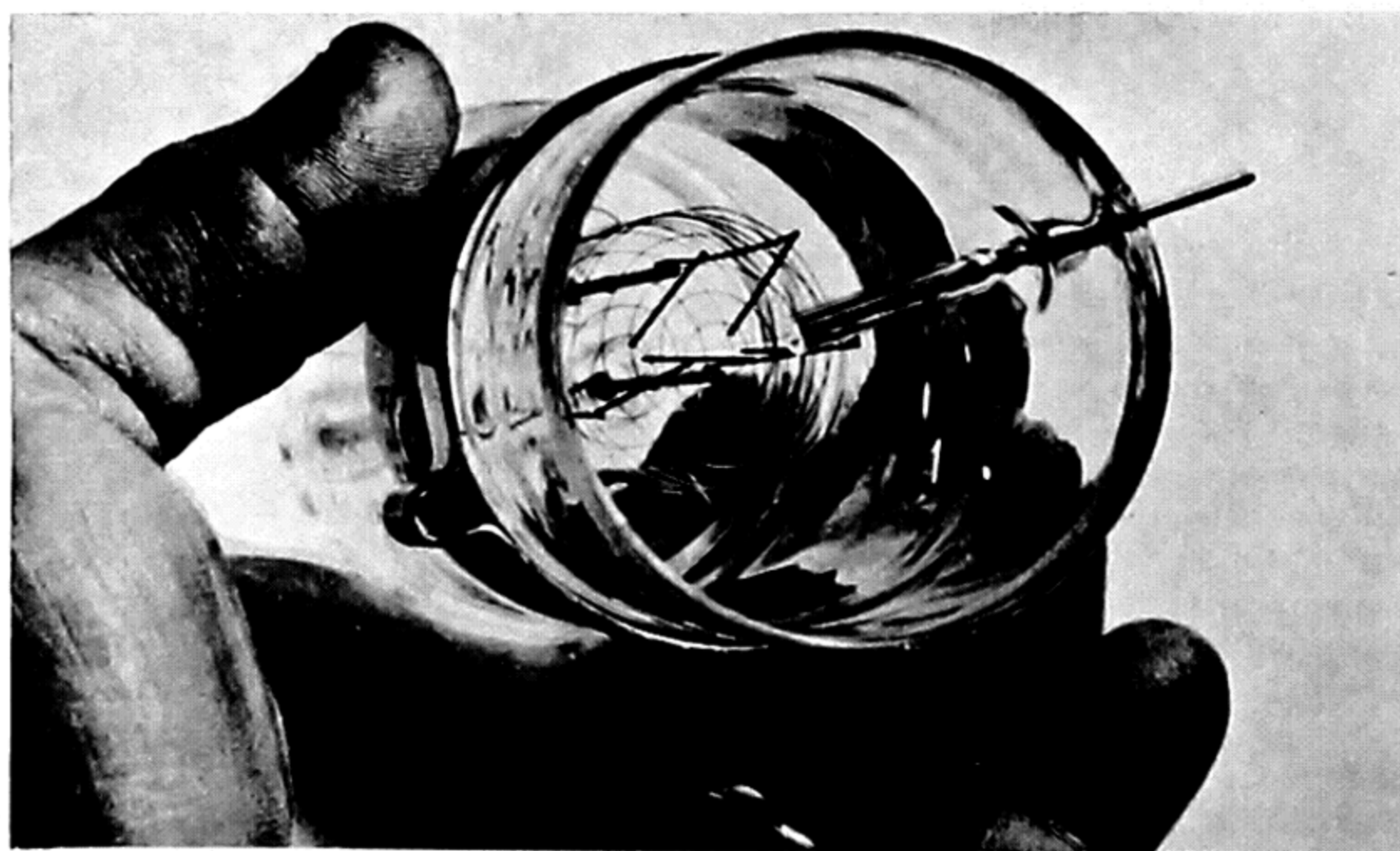
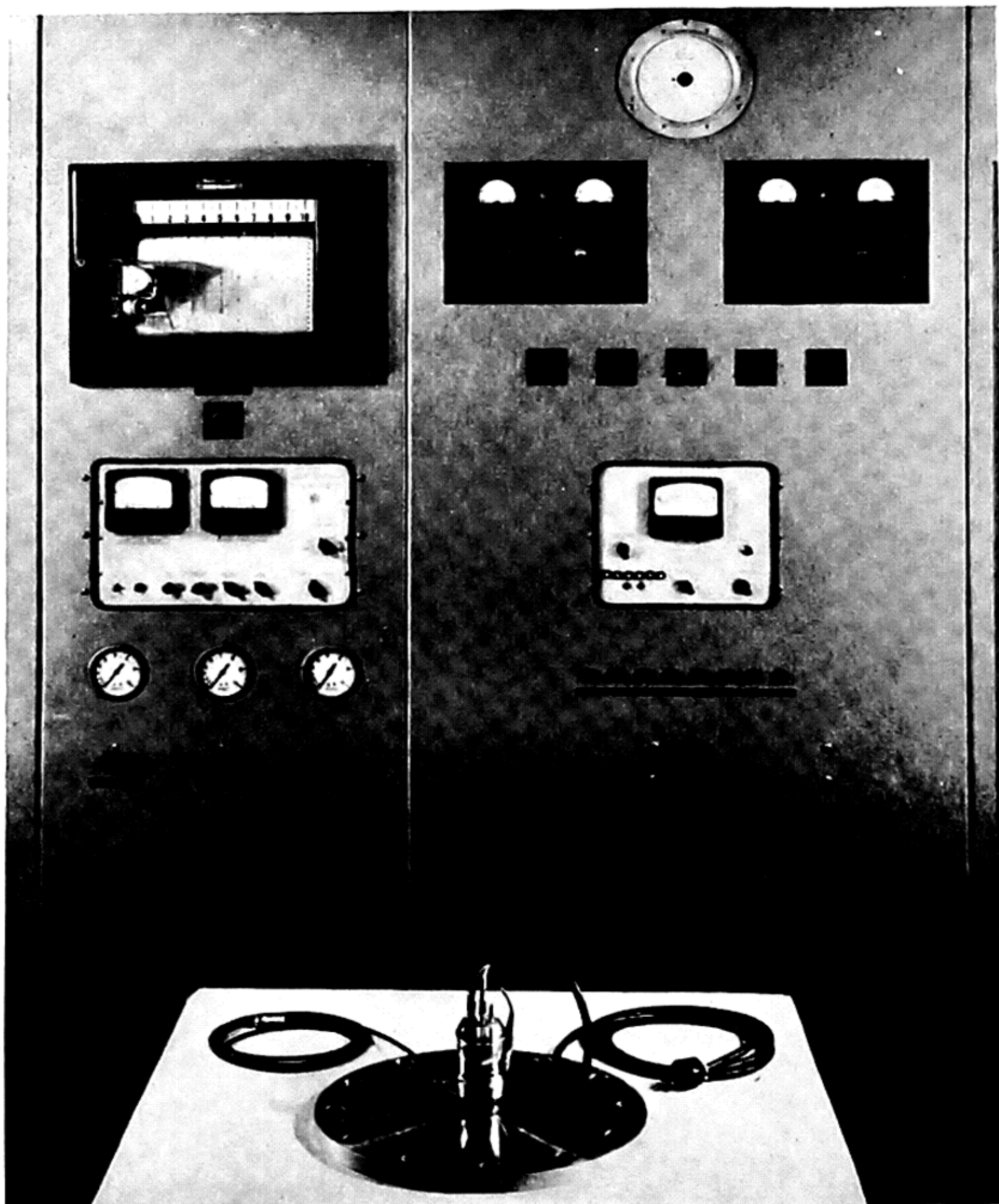
Measurements at 10^{-10} torr and below can now be made with cold-cathode gauges, in which ionizing electrons are produced by applying a high voltage to a pair of unheated electrodes [see illustration at top on page 660]. Lacking a separate grid, the cold-cathode gauge generates no spurious X-ray current. One version of the instrument developed by one of the authors (Redhead) can measure pressures as low as 10^{-12} torr.

The sensitivity of most ultrahigh-vacuum gauges depends on the composition of the gas being measured. In turn, the mixture of molecular species making up the residual gas in an ultrahigh-vacuum system depends largely on the system involved and the pumping methods. For example, a small glass apparatus exhausted with an ion pump contains chiefly helium from diffusion through the walls. On the other hand, hydrogen and hydrocarbons constitute most of the gas in a metal chamber evacuated by an oil diffusion pump. For many purposes it is more important to know the partial pressure of a particular gas or group of gases

ULTRAHIGH-VACUUM REGION

10^{-10}	10^{-11}	10^{-12}
4×10^{10}	4×10^9	4×10^8
5×10^7	5×10^8	5×10^9
3×10^6	3×10^5	3×10^4

high vacuums, density is measured in terms of the other characteristics shown here.



CONTROL PANEL (top) for pumps shown on page 655 includes a recorder (left) for ultrahigh vacuums and gauges (right) for earlier stages and subsidiary pumping systems. On the table in front is the Bayard-Alpert-type gauge to be installed in the chamber. In the close-up of the gauge (bottom) the transverse element is the ion collector, the fine spiral wire is the grid and the heavier diagonal wires are the filaments that supply electrons.

than the total pressure. What is needed is a mass spectrometer rather than a total-pressure gauge. Moreover, at very low pressures—below 10^{-12} torr—total-pressure gauges are scarcely less complex than mass spectrometers. Except in a few special cases, total-pressure ionization gauges have little to recommend them in apparatus designed for pressures below 10^{-12} torr.

Mass spectrometers, although they come in a wide variety of forms, all work in fundamentally the same way: the gas to be measured is ionized and the resulting positive ions are made to move at high speed through a magnetic, electric or combined field. Ions with different masses (but the same charge) follow different paths and are thereby distinguished from one another. The current produced by ions that have traced a particular path is an index of the number of gas molecules of a particular mass.

Spectrometers intended for ultrahigh-vacuum work must meet far stricter requirements than are demanded of ordinary analytical instruments. For one thing, they must be extremely sensitive; for another, their operation must not change the composition of the gas in the chamber. This means that they must withstand the preparatory high-temperature baking to which all ultrahigh-vacuum components are subjected.

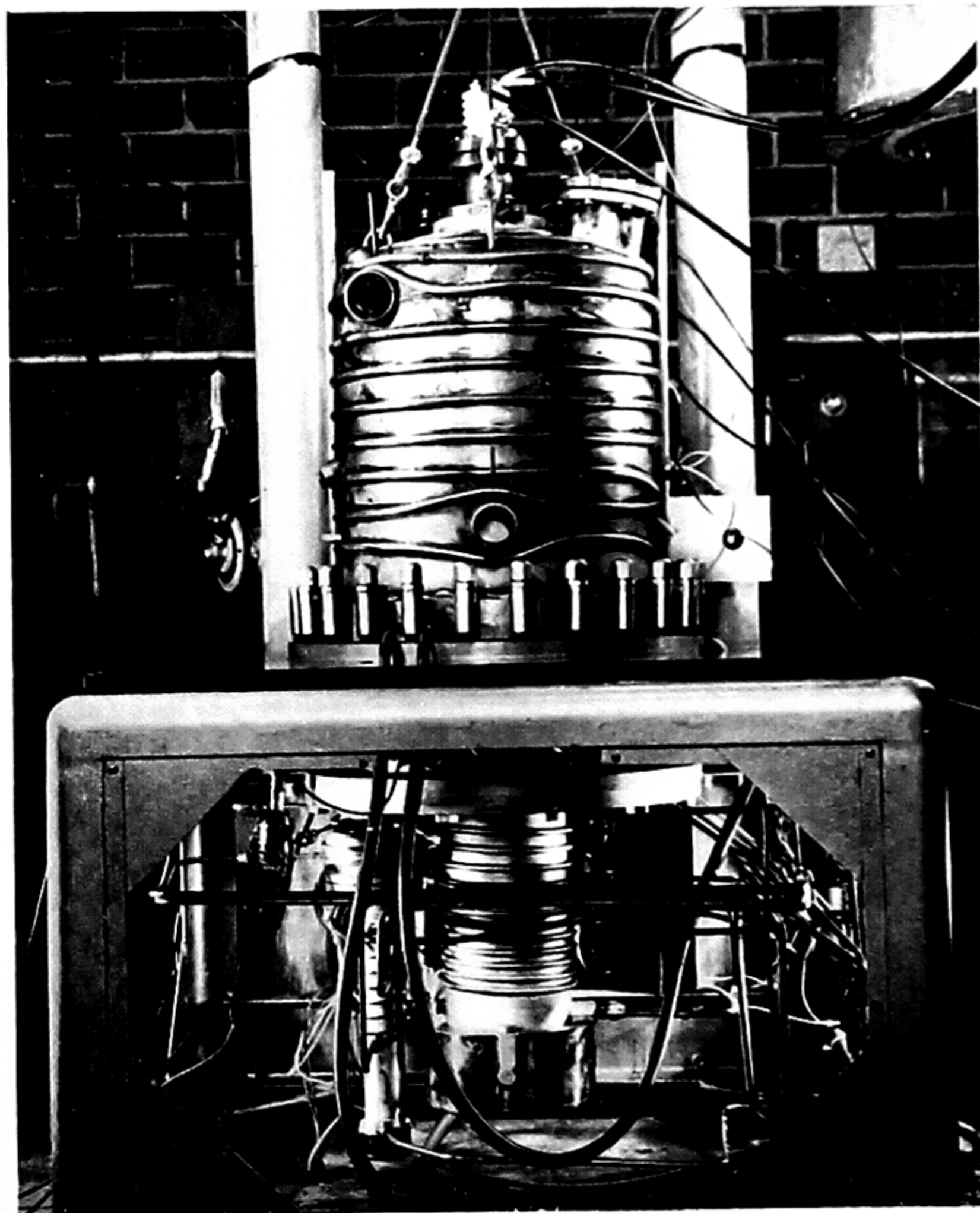
A number of quite different designs that satisfy the requirements of ultrahigh vacuum are now available. One, developed by W. D. Davis and Thomas A. Vanderslice of the General Electric Research Laboratory, is a modification of a conventional 90-degree sector instrument, in which the ion path is bent through a right angle [see top illustration on page 661]. By varying the magnetic field, ions of different mass are directed at the collecting plate, and the corresponding current is measured with the help of an electron multiplier. The instrument can measure partial pressures as low as 10^{-14} torr and can distinguish ions differing in mass by 1 per cent.

Another popular form of spectrometer is the Omegatron, which works like a cyclotron [see bottom illustration on page 660]. A radio-frequency field applied at right angles to the magnetic field whirls the ions in curved paths. At each value of the frequency, molecules of a particular mass will follow a smooth spiral to the collector plate, whereas others will be diverted. The Omegatron has a lower limit of about 10^{-11} torr. Because of the shape of the instrument and

the strong magnetic field that it requires, the addition of an electron multiplier to extend its range is most difficult.

A device called the massen filter, originally proposed by W. Paul of the University of Bonn in Germany, employs another arrangement of radio-frequency electrodes. It consists of four parallel, equally spaced circular rods carrying both direct-current and radio-frequency voltages [see bottom illustration on page 661]. The ion beam is shot down the axis of this structure. For a particular value of the voltages, ions of one mass have a stable path, oscillating closely around the axis and striking the collector plate at the far end. Ions of different masses follow trajectories that make them strike the rods. Changing the voltages changes the mass of the ions collected. With electrodes 25 centimeters long it has been possible to distinguish ions differing in mass by 1 per cent. The addition of an electron multiplier to the massen filter is relatively simple, and the combination has measured partial pressures down to 10^{-15} torr.

For the past few years ultrahigh-vacuum workers have been consolidating their position at about 10^{-12} torr. Today this pressure is obtained and measured consistently, whereas a few years ago it could be reached only by heroic efforts. Now a new cycle is beginning. Present developments in the design of gauges promise to carry them into the range of 10^{-18} torr total pressure. Improvement in pumping can be expected to follow. In fact, as has been indicated, cryogenic techniques may already be capable of producing pressures as low as those in interplanetary space.



ULTRAHIGH-VACUUM SYSTEM made by NRC Equipment Corporation of Newton, Mass., is tested. Oil diffusion pumps (*bottom*) evacuate the chamber on the table to 10^{-9} torr. A furnace (*visible at top right*) swings over the chamber to bake it before experiments.

The Authors

H. A. STEINHERZ and P. A. REDHEAD are respectively manager of the Engineering and Development departments of the NRC Equipment Corporation in Newton, Mass., and a member of the Electron Physics Group of the National Research Council of Canada. Steinhertz, who is also an instructor in vacuum technology at Boston University, acquired a B.S. in physics at Yale University in 1947. During the next four years he worked for the Westinghouse Electric Company while studying physics and mathematics at the University of Pittsburgh. From 1951 to 1955 Steinhertz was a materials engineer with the General Electric Company. He joined the NRC Equipment Corporation in 1955. Redhead, born in England in 1924, took a B.A. in physics at the University of Cambridge in 1944. He was engaged in research on proximity fuzes and microwave tubes for the British Admiralty until 1947, at which time he went to the National Research Council of Canada.

Bibliography

THE INTERPLAY OF ELECTRONICS AND VACUUM TECHNOLOGY. J. M. Lafferty and T. A. Vanderslice in *Proceedings of the IRE*, Vol. 49, No. 7, pages 1136-1154; July, 1961.

PRODUCTION AND MEASUREMENT OF ULTRAHIGH VACUUM. D. Alpert in *Handbuch der Physik*, Vol. 12, pages 609-663; 1958.

THE PRODUCTION AND MEASUREMENT OF ULTRAHIGH VACUUM (10^{-8} – 10^{-13} mm Hg). P. A. Redhead in *Fifth National Symposium on Vacuum Technology Transactions*, pages 148-152; 1959.

SCIENTIFIC FOUNDATIONS OF VACUUM TECHNIQUE. Saul Dushman. Edited by J. M. Lafferty. John Wiley & Sons, Inc., 1962.

VACUUM EQUIPMENT AND TECHNIQUES. Andrew Guthrie and Raymond K. Wakerling. McGraw-Hill Book Company, Inc., 1949.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

RADIO GALAXIES

by D. S. Heeschen

Certain galaxies are extraordinarily strong sources of radio waves. Until recently it was thought that they were galaxies in collision, but now astrophysicists seek other explanations of their radiation.

The discovery that the earth receives radio waves from space was made by Karl G. Jansky in 1931. At first it seemed that the radio waves represented a general background emission originating within the disk of our own galaxy. Much later it became evident that some of the radio energy emanated from discrete sources, either within our galaxy or outside it. Finally, in 1951, Walter Baade, using the 200-inch telescope on Palomar Mountain, made

the first identification of a discrete radio source with an optically visible object. His photographs showed that the intense cosmic source known to radio astronomers as Cygnus A coincided with the position of a galaxy—or perhaps a pair of galaxies in collision—now estimated to be 700 million light-years away. In spite of its vast distance Cygnus A emits radio waves so energetically that it had been the first “bright” region in the radio sky to be recognized as a dis-

crete source. Cygnus A was discovered in 1948 as a result of the work of J. S. Hey and his colleagues in England and J. G. Bolton and G. Stanley in Australia.

Baade's identification of Cygnus A with a remote galaxy opened new areas of astronomical investigation and introduced astrophysicists to a new type of celestial object. Both the appearance and the spectrum of the galaxy were most peculiar, and of equal pecu-



INTERACTING GALAXIES known as 4038 and 4039 in the New General Catalogue (NGC) belong to the class of “peculiar” galaxies that are strong emitters of radio waves. NGC 4038-39, which is

about 65 million light-years away, emits about 100 times more radio energy than radio galaxies classified as “normal.” This photograph was made with the 200-inch telescope on Palomar Mountain.

liarity was the fact that the galaxy was radiating more energy at radio than at light, or optical, wavelengths. The nature of the physical processes going on in this and other radio galaxies today presents one of the most fascinating and puzzling problems of astronomy.

Several thousand discrete radio sources have now been catalogued with the aid of the large radio telescopes in Cambridge, England, and Sydney, Australia. Some of the sources are associated with relatively nearby objects in our own galaxy, such as emission nebulae, composed of intensely hot interstellar gases, and other nebulae that are the remnants of supernovae. Other sources, like Cygnus A, have been identified with extragalactic objects, and there is now considerable evidence that the majority of radio sources are associated with galaxies other than our own. Sources as powerful as Cygnus A could be detected by radio telescopes even if they were far beyond the range of present optical telescopes, and presumably some of the radio sources not yet identified with visible objects are too distant to be seen.

About 100 of the discrete radio sources have now been identified with visible galaxies. These radio galaxies can be loosely divided into two groups, the "normal" and the "peculiar," according to their optical and radio characteristics. The division between the two groups is not sharp, and at present there is no unambiguous way to specify whether a radio galaxy is normal or peculiar. The terms are useful, however, and in the following paragraphs I shall try to define them by describing some observable features of the galaxies to which they apply.

All the bright nearby spiral galaxies, such as the familiar galaxies in the constellations of Andromeda and Triangulum, are weak radio sources. Most of what we know about their radio emission is due to the work of R. Hanbury Brown and C. Hazard in England and B. Y. Mills in Australia. Part of the radio emission from these galaxies is associated with the visible disk of the galaxy and part originates in an extended halo that often occupies a volume 10 or more times greater than that of the visible disk. Judging from the results of observations of the nearer galaxies, it is reasonable to expect that most of the spiral galaxies, and probably the irregular galaxies as well, are weak radio emitters. These are the normal radio galaxies.

The total energy emitted in the radio-frequency region of the spectrum by a normal radio galaxy is on the order of 10^{38} ergs per second, or, to use a more

familiar measure of radio power, 10^{28} kilowatts. This is about 1,000 times the radio output of the most intense radio sources within our own galaxy, but it is still only about a millionth of the energy that galaxies emit at optical wavelengths.

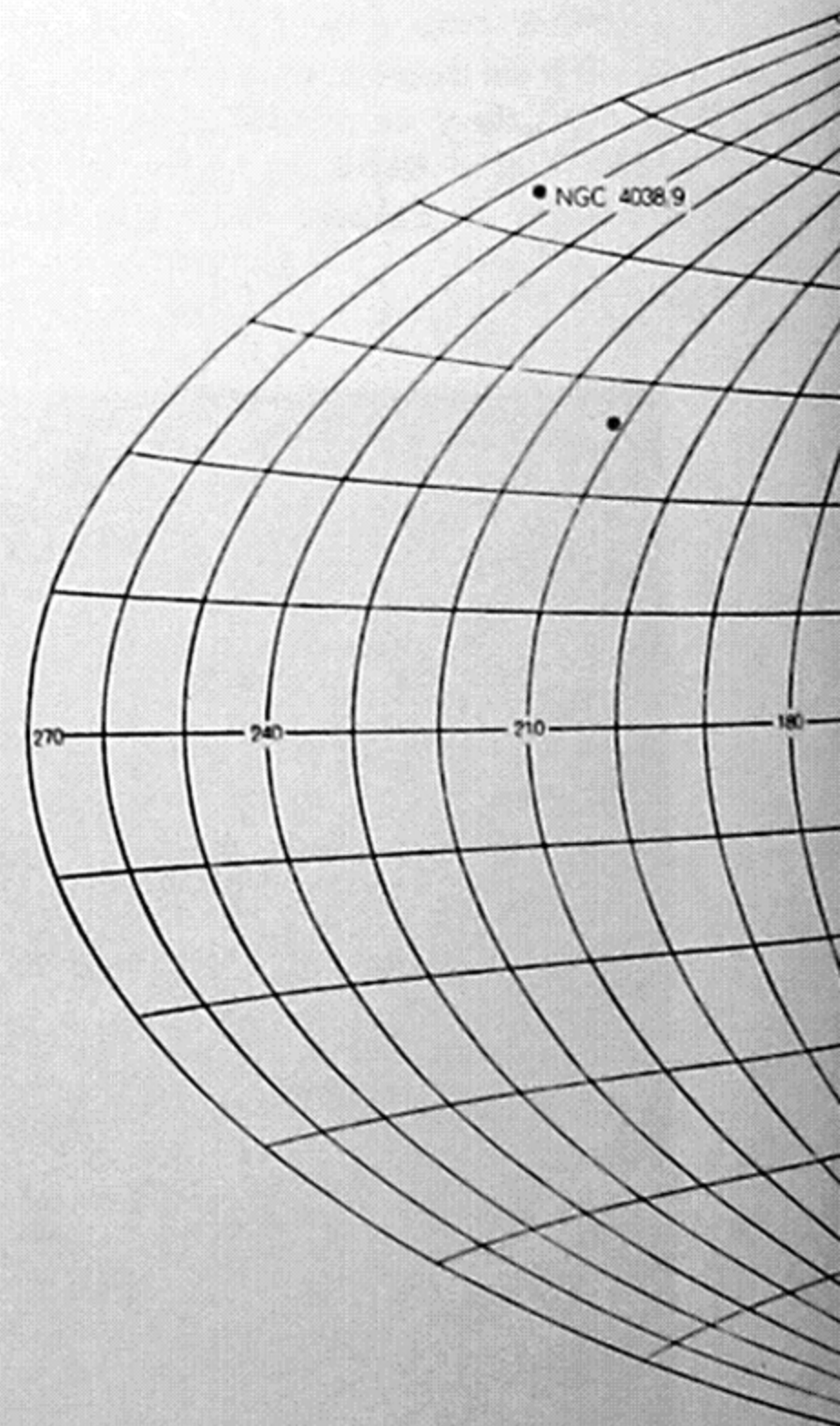
Some galaxies of standard appearance emit significantly more than 10^{28} kilowatts of radio power. For example, the spiral galaxy NGC 1068 emits about 10^{30} kilowatts of radio energy, or about 100 times more energy than the radio galaxies classified as normal. Moreover, the radio emission is concentrated in a source of small diameter at the center of the galaxy. If there is any emission from an extensive halo, it is too weak to be detected with present instruments. The optical spectrum of NGC 1068 shows intense, broad emission lines that suggest a high degree of energetic, chaotic activity in the nucleus of the galaxy. It is tempting to try to connect this phenomenon with the enhanced radio emission, but several other spiral galaxies whose optical spectra exhibit the same features do not show enhanced radio emission.

NGC 1068 is one of the weaker of the galaxies that have come to be known as peculiar radio galaxies. They are peculiar in that they emit radio waves much more intensely than do the normal spiral galaxies. Although many of them also have visual or spectral peculiarities of one sort or another, this is not invariably the case. The well-known elliptical galaxy M 87 is another example of a peculiar radio galaxy. The radio emission from M 87 is almost 100 times more intense than that from NGC 1068. The emission appears to come from two concentric sources, approximately centered on the visible galaxy: a small and intense core, and a large and less intense halo. On long-exposure photographs M 87 appears as a normal giant elliptical galaxy. Short-exposure photographs, however, reveal the presence of a bright jet extending from the galactic center [see illustration on page 672]. The jet is suggestive of material being ejected with high energy from the nuclear region of the galaxy. Observations made with the 200-inch telescope by Baade have shown that the light radiated from the jet is strongly polarized. As we shall see, this observation has important consequences for the interpretation of radio emission from extragalactic radio sources.

A number of other intense radio sources have been identified with elliptical galaxies that have a normal appearance. On the other hand, some comparatively close giant elliptical galaxies

have no detectable radio emission. It now appears that only a small fraction of normal-looking elliptical galaxies emit enough radio energy to be called peculiar. Whether or not the peculiar ones also have jets similar to that seen in M 87 is not known, since the jet of M 87 would not be easily detectable at the distances of other elliptical galaxies that are strong radio sources. It seems likely, however, that the phenomenon responsible for the jet in M 87 is related to the cause of the intense radio emission from this object.

Many radio sources have been identified with galaxies that cannot easily be classified by the usual scheme, which is based solely on form. The galaxy associated with the source Cygnus A, for example, appears visually as two nuclei in contact, surrounded by a faint common envelope [see illustration on page 671]. This has been interpreted variously as two galaxies in collision, a single galaxy whose nucleus is splitting



MAP OF RADIO GALAXIES shows the location, in galactic co-ordinates, of 45 of the

apart and a single galaxy with a peculiar lane of dust across its central region. The radio emission from Cygnus A is about 10^{34} kilowatts, a million times more than that from a normal radio galaxy. NGC 5128 is another example of a radio galaxy whose visual features cannot readily be interpreted [see bottom illustration on page 672]. It looks like an elliptical galaxy with a band of dust running through its center. Several other radio galaxies are now known whose appearance is similar to that of NGC 5128 and Cygnus A.

A striking feature of these powerful radio sources is the spatial distribution of their radio emission. In Cygnus A the radio emission originates not in the region of the visible object but in two regions symmetrically placed on each side of the galaxy at distances of about 100,000 light-years! NGC 5128 has four regions of radio brightness, two resembling those in Cygnus A and two lying

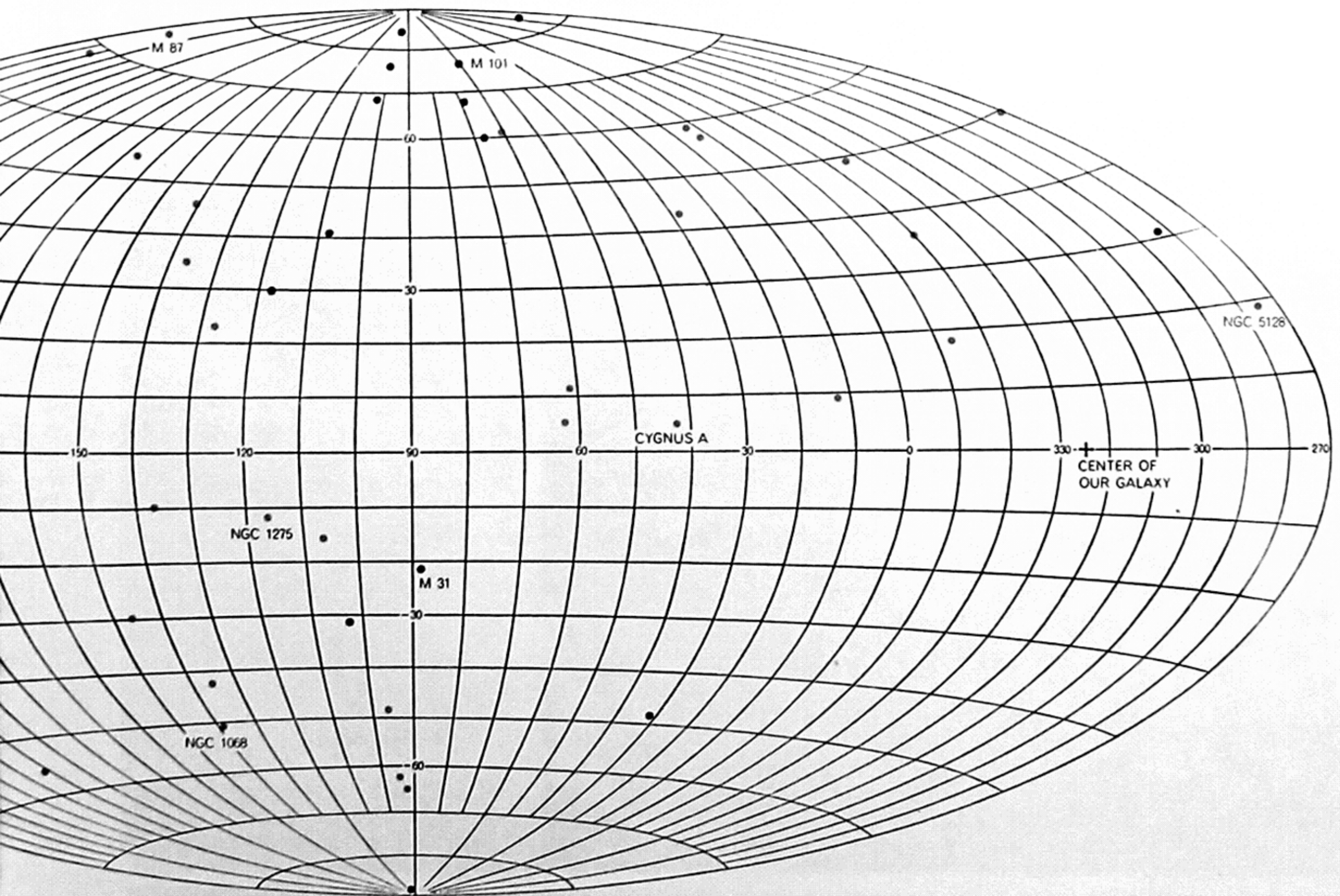
closer to the central portion of the galaxy. All four emissive regions lie approximately on a line running through the center of the galaxy and perpendicular to the dark absorbing band. Recent observations at the California Institute of Technology and at Jodrell Bank in England have shown that many radio galaxies are similar to Cygnus A in that they have two radio-emitting regions of approximately equal intensity. Whether or not some of these radio galaxies may be more complex, like NGC 5128, cannot at present be determined. With available equipment the complex brightness distribution of NGC 5128 can be observed only in nearby sources.

To summarize, there seem to be at least three different types of peculiar radio galaxy: normal-appearing spiral galaxies with enhanced radio emission, like NGC 1068; normal-appearing elliptical galaxies, perhaps with central jets like M 87, or with some other peculiar-

ity; galaxies with a complex radio brightness distribution and a peculiar unclassifiable visual appearance, like that of NGC 5128 or Cygnus A. The intensity of radio emission from the last two types may range from only slightly more intense than a normal galaxy (10^{28} kilowatts) all the way up to 10^{34} kilowatts, a range of more than 10^5 .

To this list of types of radio galaxy we might also add objects such as the twin galaxies NGC 4038-39. Photographs suggest that this close pair of galaxies is in strong gravitational interaction. The pair has a weak radio output, of an intensity comparable to that from NGC 1068. Several other pairs of interacting objects of this general appearance have also been identified with radio sources.

This classification of radio galaxies into four general categories may or may not be meaningful. It is not at all certain that the four categories represent four



most carefully studied radio sources lying outside our galaxy. The normal radio galaxies are indicated by black dots; peculiar galax-

ies, which emit from 100 to approximately a million times more radio energy than do normal galaxies, are shown by colored dots.

distinct types of radio source. Nor is it clear that all radio galaxies can properly be placed in one of the four categories. We must keep in mind that barely 100 radio sources have so far been identified with galaxies, and almost every one appears unique in some sense. Thus it is quite possible that the physically important unifying features have not yet been recognized. The above groupings do, however, serve to classify the major observed characteristics of peculiar radio galaxies, whether or not the groupings have physical significance.

The radio energy emitted by most galactic and extragalactic radio sources increases in intensity with increasing wavelength. This is described as a nonthermal distribution of energy because it is just the opposite of the distribution expected from a body of hot gas, such as a star. In a hot gas inter-

actions between electrons and protons arise as a result of random thermal motions, and these motions produce radiation in both the optical and radio regions of the electromagnetic spectrum. The intensity of this thermal emission remains essentially constant over a wide range of wavelengths, but beyond a certain point the intensity decreases as the wavelength becomes greater. Only the radio sources associated with the emission nebulae of our own galaxy exhibit this typical thermal emission spectrum. All other discrete radio sources have a nonthermal spectral distribution. The mechanism for producing such a spectral distribution remained for many years one of the major puzzles of radio astronomy.

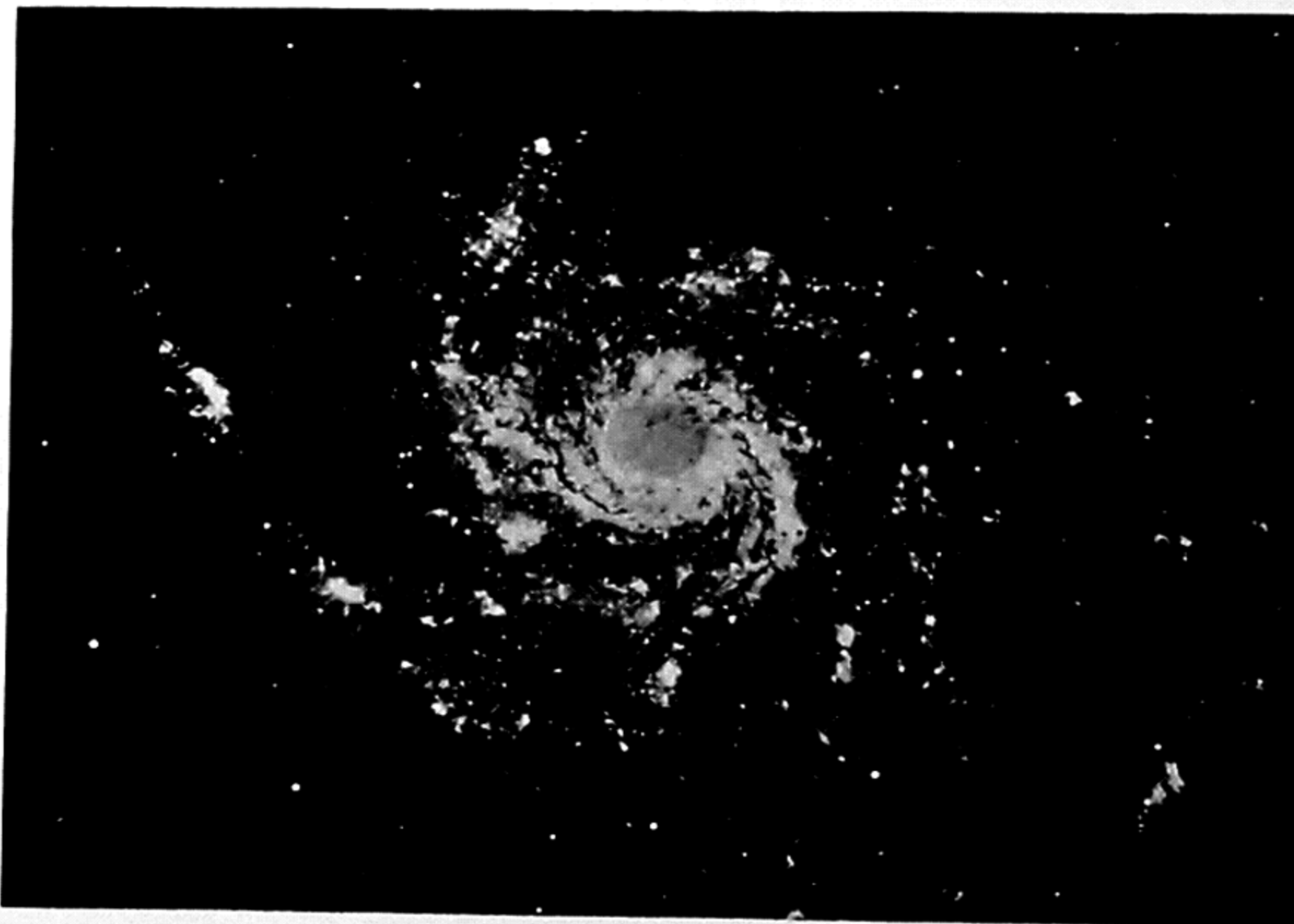
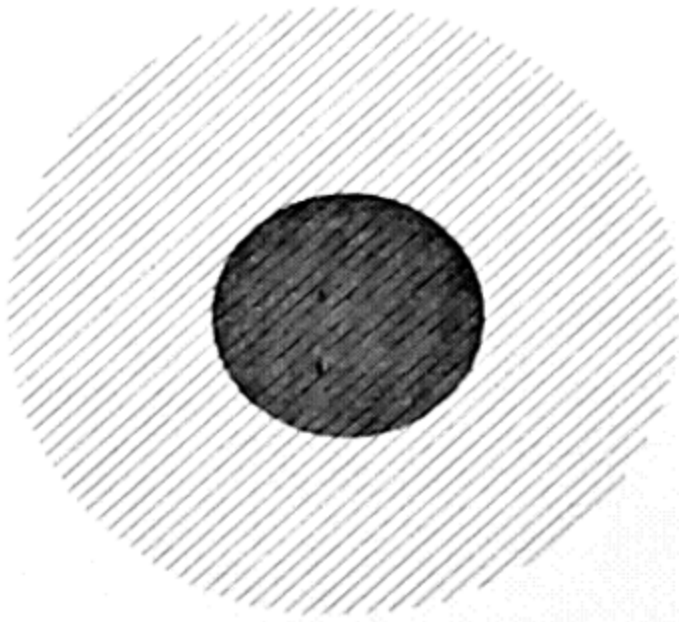
The first plausible mechanism was developed in 1953 by the Soviet astrophysicist I. S. Shklovsky. He theorized that the radio emission and much of the optical emission from the Crab nebula, the remnant of a supernova and a strong radio source in our own galaxy, was produced by the interaction of energetic electrons and a magnetic field associated with the nebula. A relativistic electron (that is, an electron moving with a speed close to that of light) spiraling around a line of magnetic force is known to emit intense radiation in a narrow band of wavelengths determined by the energy of the radiating electron and the strength of the magnetic field [see illustration on page 674]. This radiation is some-

times called synchrotron radiation, because it is produced when electrons are accelerated in the magnetic field of a synchrotron. In the weak magnetic fields found in interstellar space, electrons with energies in the range of one billion to 10 billion electron volts radiate at radio wavelengths, whereas electrons of higher energy radiate at optical wavelengths. The spectral distribution of the radiation produced by this mechanism is similar to the energy distribution of the electrons. Since the known energy distribution of cosmic rays is similar to the radio-frequency spectral distribution of the Crab nebula, Shklovsky concluded that relativistic electrons with essentially the same energy distribution as that of cosmic rays were responsible for the optical and radio emission from the Crab nebula. He also predicted that the optical emission from the Crab nebula would show a high degree of polarization, because this is another characteristic of the synchrotron radiation.

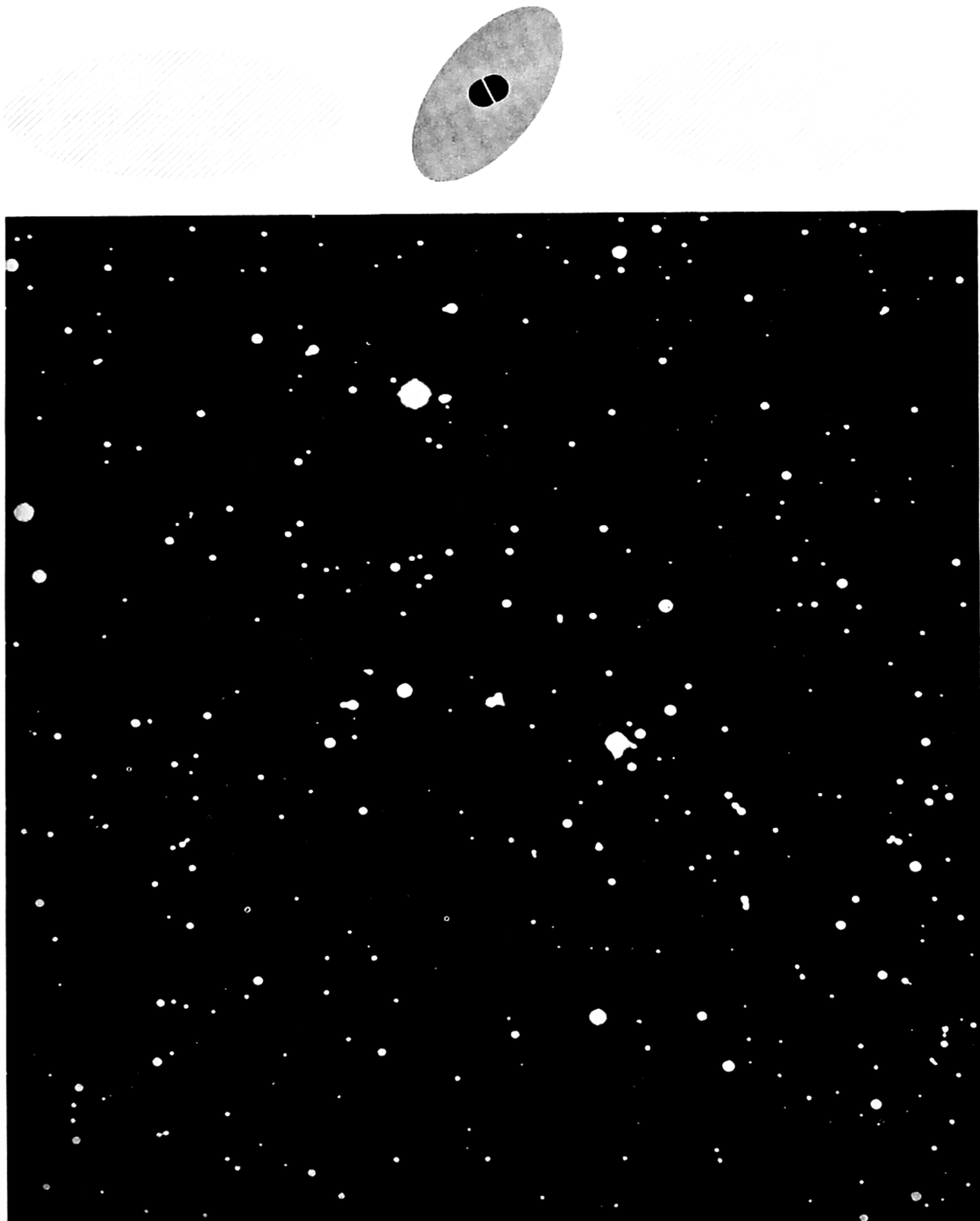
The prediction was beautifully confirmed by the Soviet astronomers V. A. Dombrovsky and M. A. Vachakidze and the Dutch astronomers Jan H. Oort and Theodore Walraven. Light from the Crab nebula was indeed highly polarized. Oort and Walraven also made a detailed investigation of the radio and optical emission characteristics of the Crab nebula and showed that many of the hitherto puzzling features of its radiation could be easily explained by the synchrotron mechanism.

These findings led naturally to the idea that the synchrotron mechanism may also account for the radiation from radio galaxies. So far, however, the only direct evidence indicating that this mechanism may be at work in radio galaxies is the important observation by Baade of the polarization of light from the jet of M 87. Although the radio spectra of radio galaxies generally show an even steeper rise in intensity with increasing wavelength than that of the Crab nebula, the spectra are still similar to the energy spectrum of cosmic rays. The synchrotron mechanism is the only one proposed so far that offers a reasonable explanation of the strongly nonthermal distribution of radio energy from extragalactic sources.

All discrete radio sources, both galactic and extragalactic, emit a fairly smooth continuum of radio-frequency energy; that is, the intensity of emission varies smoothly with frequency. The only observable absorption or emission line is the well-known line of the un-ionized hydrogen atom, which occurs in a

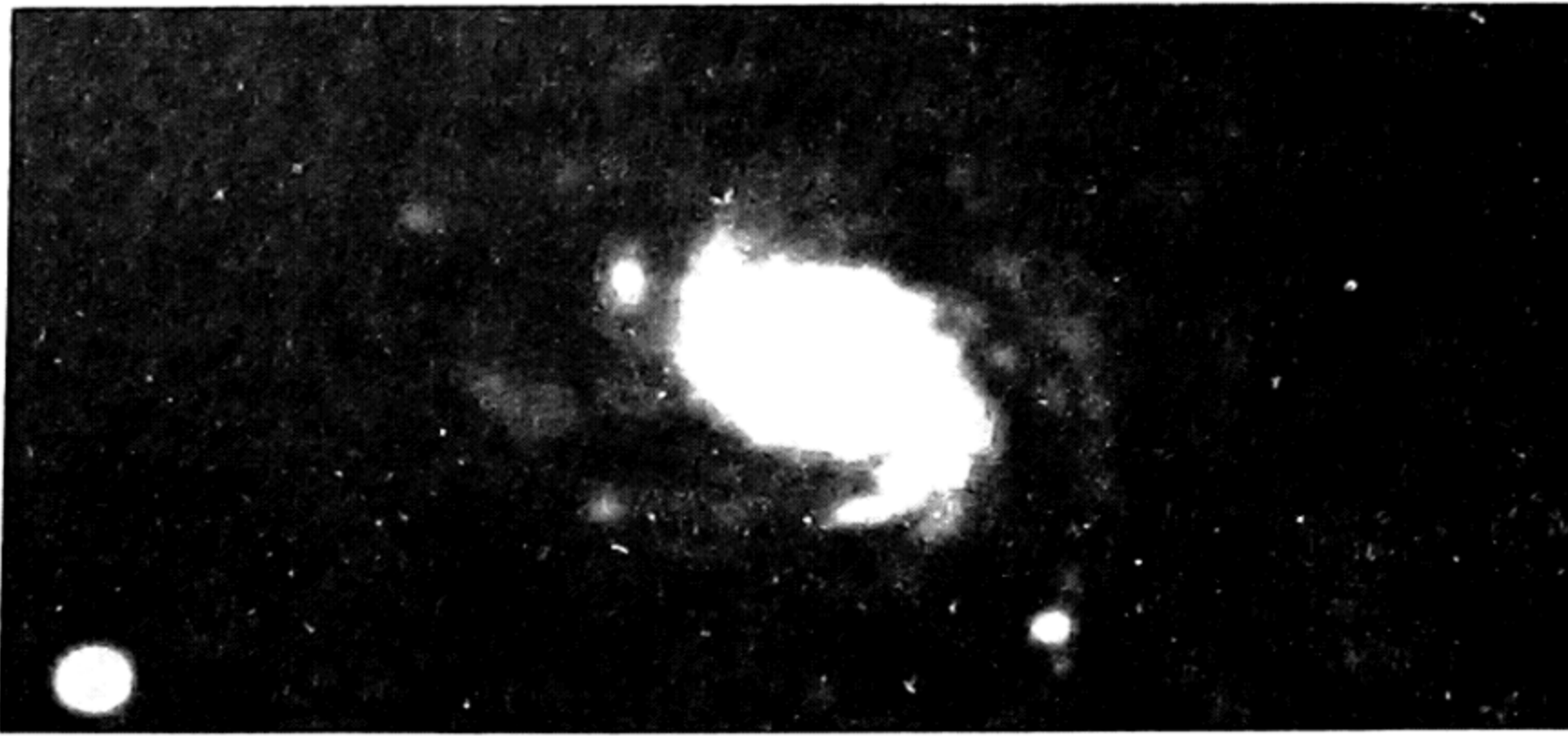


NORMAL RADIO GALAXY NGC 5457 (also known as M 101) is a well-developed spiral galaxy, about eight million light-years away. The photograph was made with the 200-inch telescope. In the diagram above the photograph the gray area represents the visible portion of the galaxy. The colored area shows the much larger halo from which radio waves emanate.

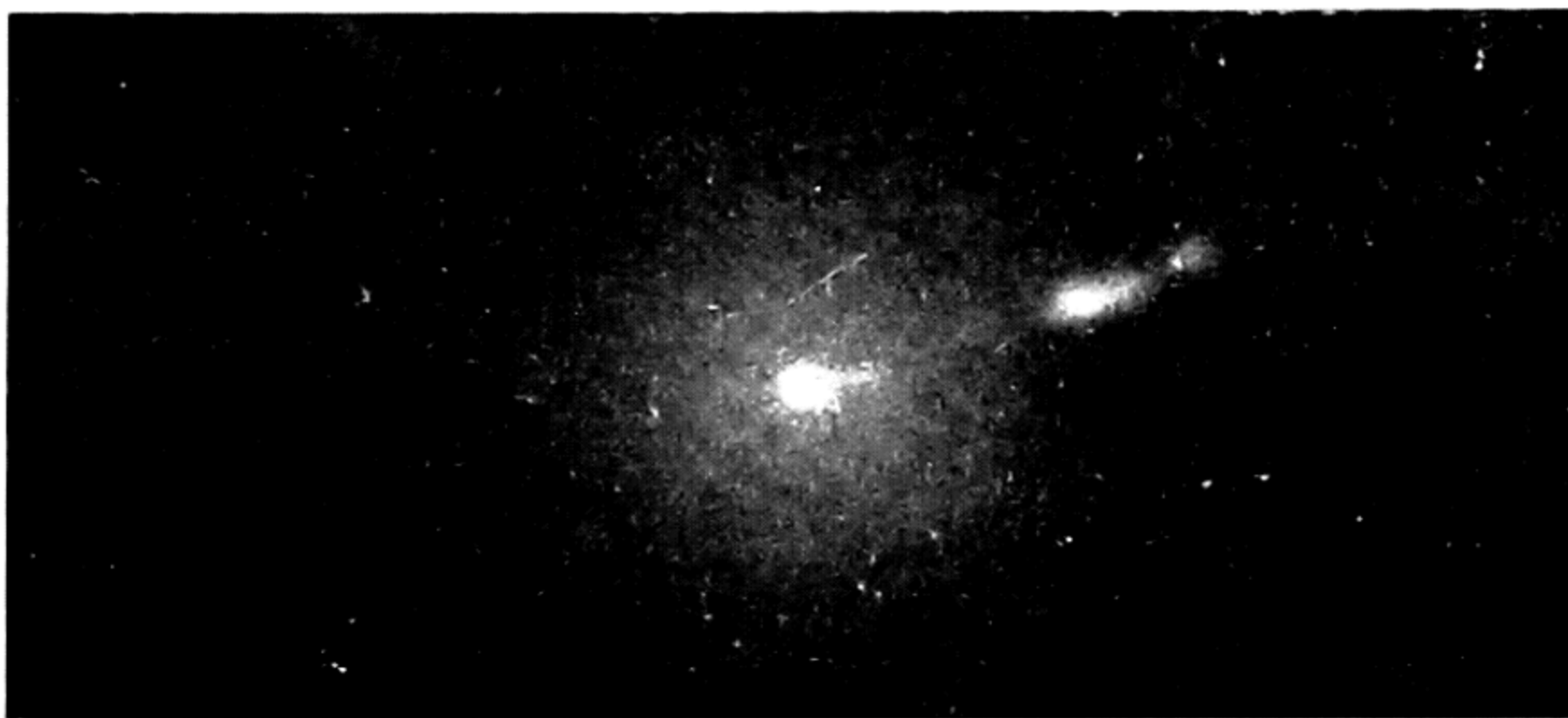


STRONGEST EXTRAGALACTIC RADIO SOURCE is the peculiar galaxy, or pair of galaxies, known as Cygnus A. In this photograph, made with the 200-inch telescope by Walter Baade, Cygnus A is the object in the center that looks like two galactic nuclei in

contact. The radio emission associated with Cygnus A is shown in color in the diagram above the photograph. The radio-emitting regions lie some 100,000 light-years to each side of the visible nuclei. Cygnus A itself is about 700 million light-years from the earth.



NGC 1068, a late spiral 40 million light-years away, emits 10^{30} kilowatts of radio energy, which makes it peculiar. The photograph was made with 60-inch telescope on Mount Wilson.



M 87, 40 million light-years away, emits 10^{31} kilowatts of radio energy. A luminous jet extends from its center. Photograph was made with the 120-inch telescope at Lick Observatory.



NGC 5128, 15 million light-years away, emits 10^{31} kilowatts of radio energy. The dark absorbing band running through the galaxy is presumably dust. Photograph was made with 200-inch telescope. Radio appearance of these three galaxies is depicted on opposite page.

narrow band of wavelengths centered at about 21 centimeters. Over all the rest of the observable radio band—from about one centimeter to 10 meters—the spectrum is continuous.

If one plots the intensity of the radio emission (expressed logarithmically) against wavelength, one obtains a sloping line that is nearly straight. The slope, which reflects how rapidly the intensity changes with wavelength, is called the spectral index. A spectral index of -1 , for example, indicates that the intensity is proportional to wavelength, whereas a spectral index of -2 means that intensity increases as the square of the wavelength. The indices carry a minus sign because in the usual way of plotting them the curves slope to the left [see illustration at bottom on page 674].

The spectral characteristics of the radio emission from radio galaxies can provide clues as to the nature of the galaxies. The spectra of all known radio galaxies, both normal and peculiar, are essentially similar. The extreme range of their spectral indices is from about -0.2 to -1.2 , but the majority have spectra with indices in the range -0.6 to -1.2 . In many cases the spectral index of a radio galaxy is approximately constant over the entire observed radio-frequency range. In other cases, however, the plotted index takes the form of a bowed curve, being steepest at the short-wavelength end of the spectrum and less steep as the wavelength increases. The spectral index of Cygnus A varies in this way from about -1.2 for wavelengths in the neighborhood of three centimeters to about -0.5 for wavelengths of 600 centimeters.

It is quite remarkable to find that the spectra of radio galaxies are so similar in view of the great diversity in their optical appearance. The basic similarity of their radio spectra has two implications. First, the radiation mechanism is probably the same in all radio galaxies. It is unlikely that different radiation mechanisms in different sources would give rise to such similar spectra. And presumably this radiation process is the synchrotron mechanism.

Second, there must be a continuing supply of relativistic electrons to replace those whose energy is dissipated through the radiation process. If this were not so, we would expect to find a wide range of spectral indices just because of this energy loss. The rate at which a relativistic electron loses energy by the synchrotron mechanism is proportional to the square of its energy, and the wavelength at which most of this energy is radiated is inversely proportional to the

square of the electron energy. Thus electrons responsible for the radiation at short radio wavelengths lose energy much more rapidly than those responsible for radiation at longer wavelengths. If at some stage in the life of a radio galaxy the supply of relativistic electrons were abruptly cut off, the short-wavelength emission would decline more rapidly than the radiation at longer wavelengths. This would lead, in a short period of time, to a spectral index that is very steep at short wavelengths. That no such steep spectra have been observed implies a continuing supply of electrons, for periods long compared to the decay time of the short-wavelength emission.

This does not necessarily mean that intense radio emission from radio galaxies is a long-lived phenomenon. The lifetime of an electron radiating at a wavelength of 10 centimeters is only of the order of 100,000 to a million years, depending on the magnetic-field strength. If relativistic electrons are supplied to the source for 10 million years, the resulting spectrum would conform to those we observe.

The change in spectral index with wavelength observed in the case of Cygnus A and several other sources finds a ready explanation in terms of the synchrotron mechanism if relativistic electrons are being continuously produced. At an early stage in the radio life of a galaxy the intensity of emission at a particular wavelength will be proportional to the total number of electrons with the appropriate energy for radiating at that wavelength. As time goes on and the supply of relativistic electrons increases, the intensity of emission at a particular wavelength will increase in proportion. The electrons are also losing energy through radiation, however. At some time, after a sufficient number of electrons have been produced, the rate of loss of energy by radiation will just compensate for the additional energy being supplied by production of relativistic electrons. At this time the intensity of emission will stop increasing and will remain approximately constant so long as electrons continue to be produced to replace the energy lost by radiation. Now the time required to reach this equilibrium condition increases with increasing wavelength. Therefore if we consider the entire spectrum of wavelengths, we will find an initial period in which the intensity at all wavelengths increases steadily so long as more and more electrons are being injected into the system. After some time, however, equilibrium will be reached at short wavelengths, and from then on the output intensity at

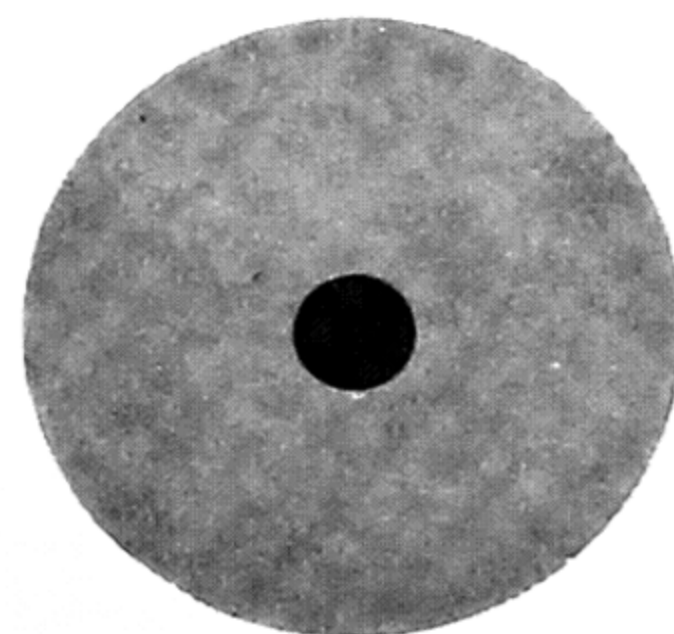
these wavelengths will not increase, whereas at longer wavelengths the intensity will continue to increase. Thus the short-wavelength portion of the spectrum becomes steeper than the long-wavelength part. Over a period of time the equilibrium point, where the spectrum changes from that of the initial injection spectrum to that of equilibrium, moves further and further to longer wavelengths. It turns out that the difference in slope between the initial injection spectrum and the steeper equilibrium spectrum is just .5. This is approximately the change in slope that is observed in some radio galaxies with spectra that steepen at shorter wavelengths.

If the magnetic field of a source can be determined in some fashion, the wavelength of maximum change of slope can be used to determine the approximate age of the source. Some estimate for the strength of the magnetic field in Cygnus A can be obtained by assuming roughly equal division of energy between the magnetic field and either the relativistic electrons or the primary cosmic rays. These assumptions yield a magnetic-field strength on the order of 10^{-4} gauss. By relating this figure to the spectrum of Cygnus A one obtains an age for this radio source of only 400,000 years. If the basic assumptions are even approximately correct, Cygnus A has been a peculiar radio galaxy for a brief period indeed as astronomers reckon time.

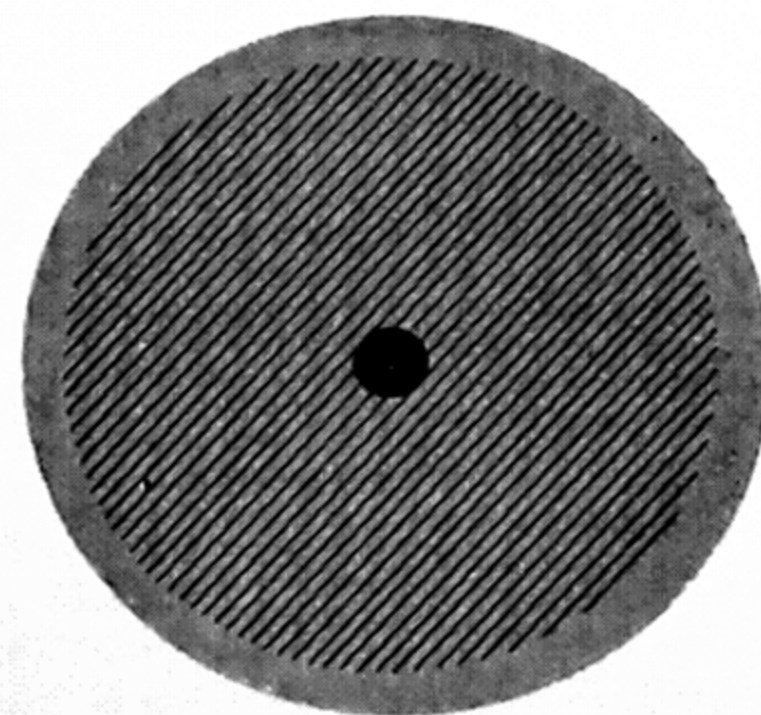
The difference of .5 between an initial injection spectrum and the final equilibrium spectrum is also interestingly close to the range in spectral indices shown by the majority of radio galaxies. Perhaps most sources have the same injection spectrum, with the differences in spectral index from one source to another then being due to differences in evolutionary age.

A number of theories have been proposed to explain radio galaxies. Normal radio galaxies can be understood fairly well in terms of known phenomena. Several investigators have shown that the cosmic ray flux in our own galaxy is sufficient to supply, by collisions with the atoms of the interstellar gas, the relativistic electrons needed to

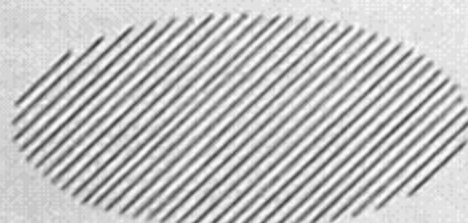
account for the observed intensity of radio emission. This is probably the case in other normal galaxies. According to this theory relativistic electrons are produced in the disk of a galaxy, where the interstellar gas is concentrated. Once produced, however, they cannot be confined to the disk and diffuse out into a halo two or three times larger than the disk. The dimensions of the halo are determined by the galactic magnetic field, which controls the distribution of relativistic electrons. Cosmic rays provide a continuing source of electrons for a long time, since the probable lifetime of a galactic cosmic ray proton is at least a billion years. On the basis of this picture a normal galaxy should emit radio energy for a long time, provided that cosmic rays do not escape from the galaxy in any appreciable number. If this notion is



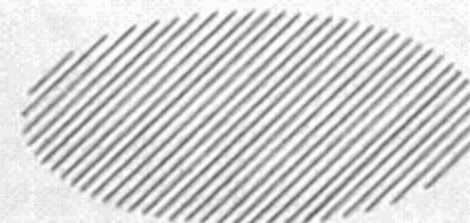
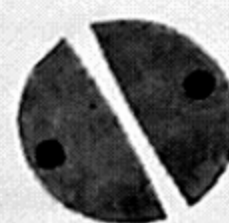
NGC 1068 emits most of its radio energy from a small central source (*color*). Gray area represents optically visible galaxy.

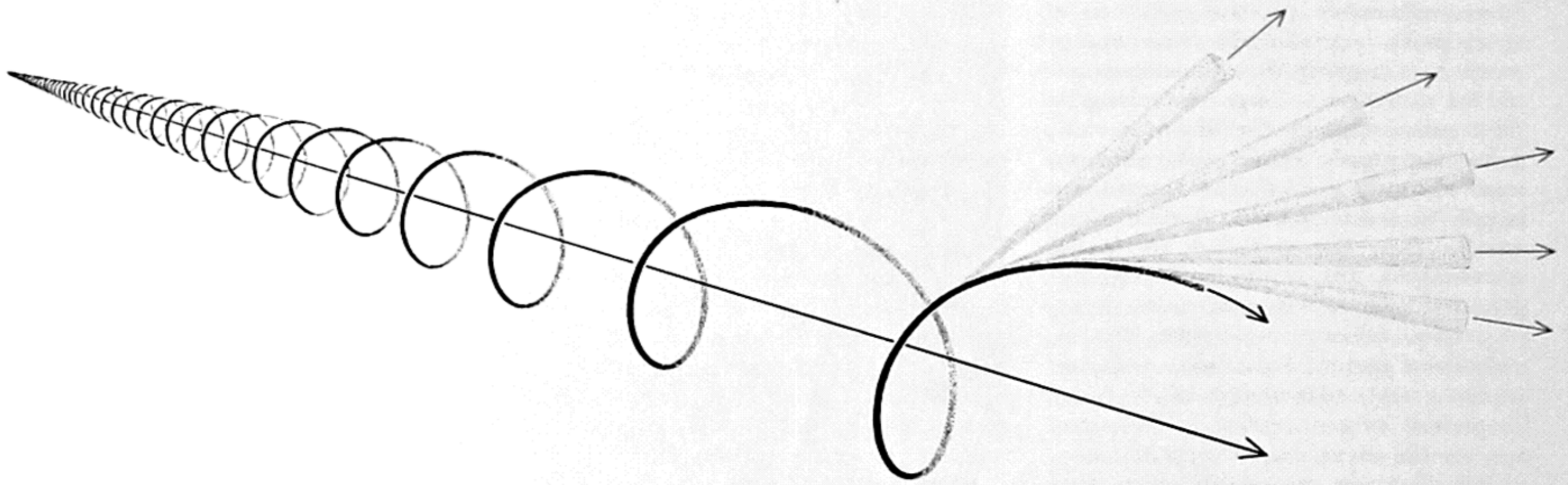


M 87 emits radio energy from two concentric sources, an intense core and a weaker halo slightly smaller than the visible galaxy.



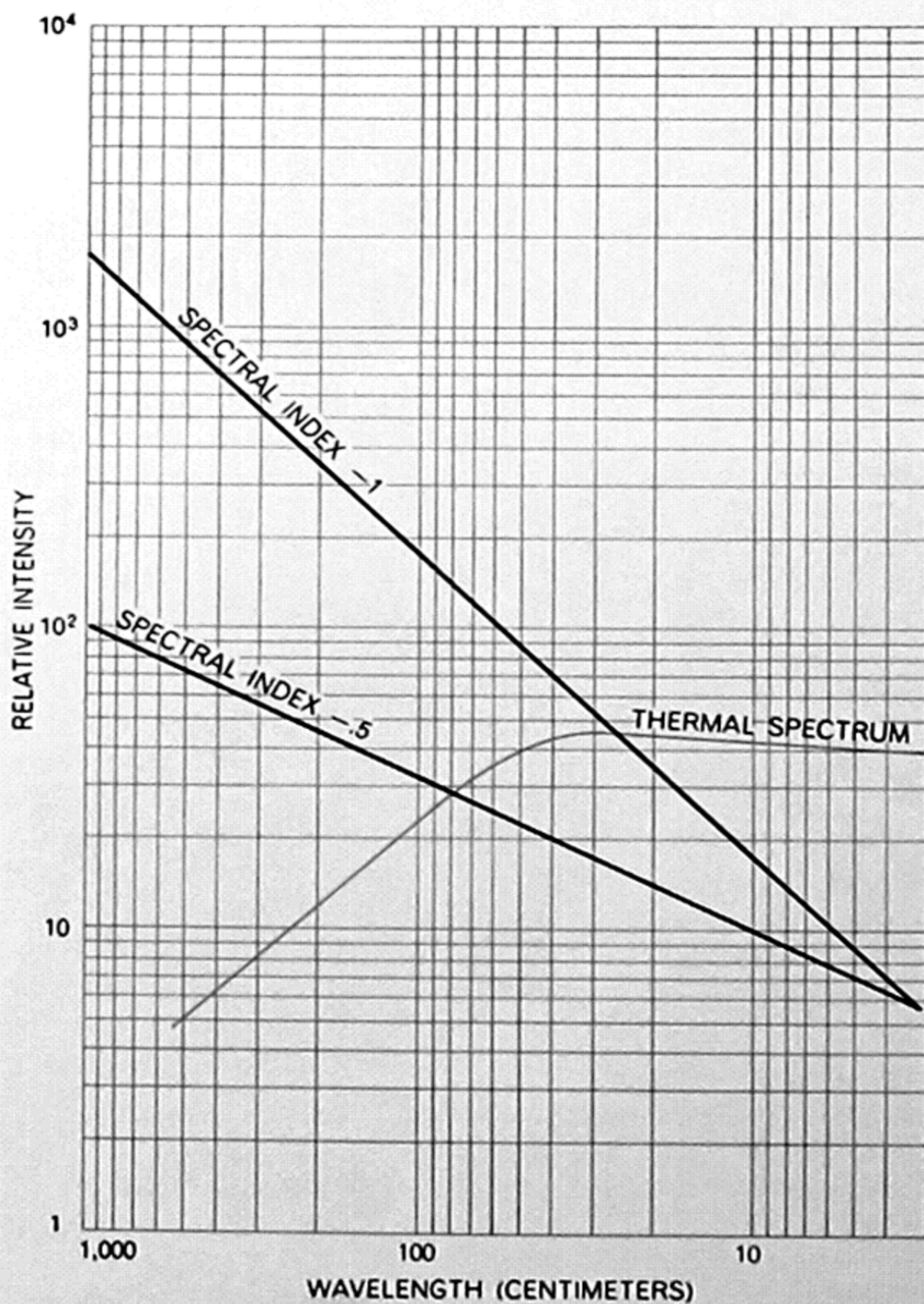
NGC 5128 shows four regions of radio brightness, two small and intense, that lie roughly on a line running through center of galaxy and perpendicular to the dust band.



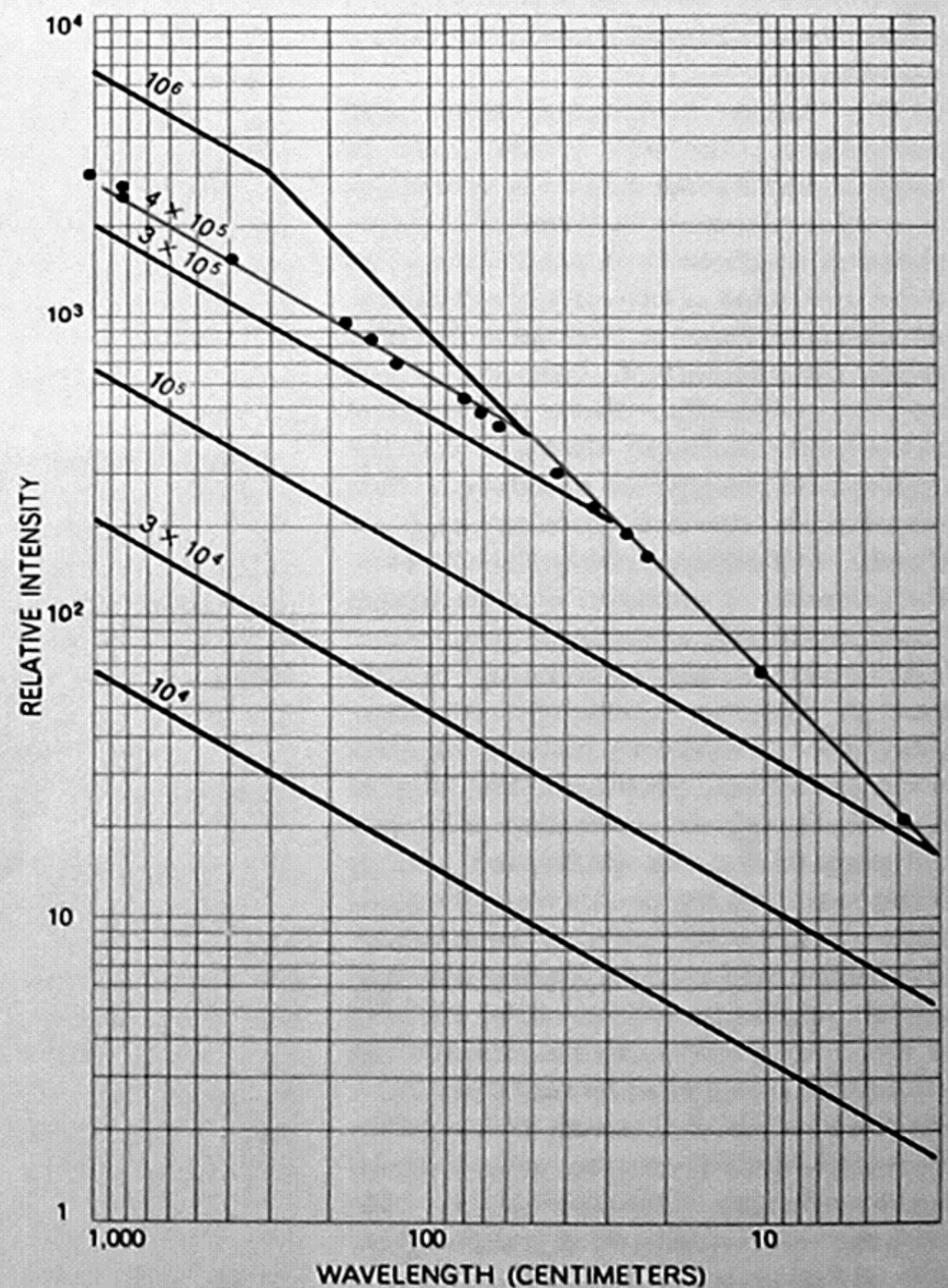


SYNCHROTRON MECHANISM has been proposed as the source of most radio energy detected by radio telescopes. When a high-speed electron spirals around a line of magnetic force, it con-

tinuously radiates a cone of energy in the direction of its motion. Under suitable conditions of electron energy and magnetic-field strength the radiation falls in the radio region of the spectrum.



SPECTRA OF RADIO GALAXIES are generally "nonthermal." This means that the intensity of emission increases with wavelength, in contrast with a typical thermal spectrum (*color*), in which the intensity falls sharply beyond a certain wavelength. Spectra of radio galaxies usually lie between the two black lines.



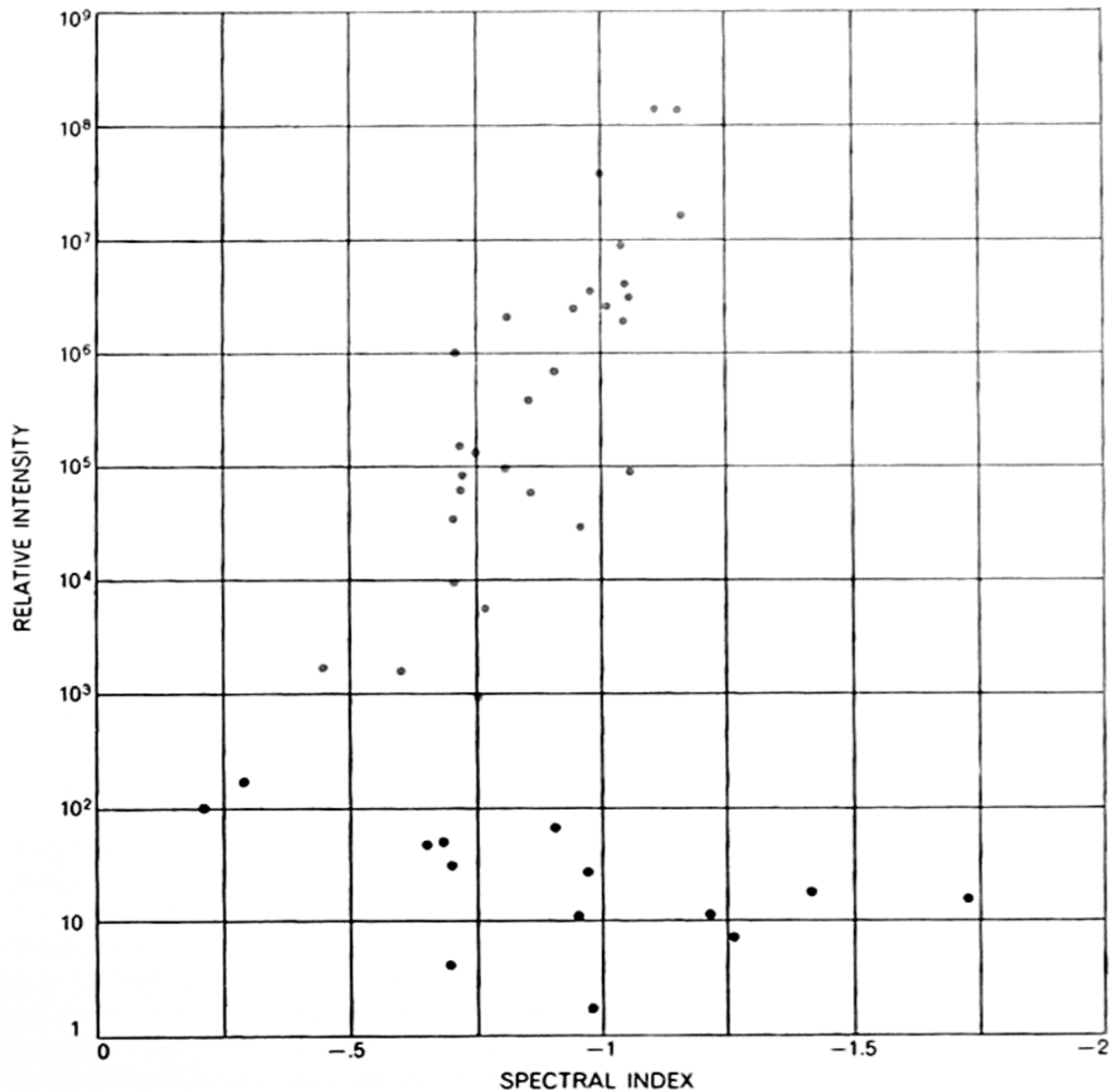
AGE OF CYGNUS A as a radio source can be estimated by computing a series of curves showing how the spectral index would change with time (*figures on curves represent years*). The dots show the observed intensity of Cygnus A at various wavelengths and indicate an age slightly greater than 4×10^5 (400,000) years.

correct, the source of cosmic rays is the remaining big unknown.

Explaining the peculiar radio galaxies is more difficult. One of the earliest suggestions was that the radio emission from Cygnus A and other intense radio sources is somehow produced as a result of the collision of two galaxies. The visual appearance of Cygnus A suggests that it may be two galaxies in collision, and the optical spectrum shows broad emission lines of highly excited atoms, suggestive of a turbulent, high-temperature gas, which might result from such a collision. Although the collision hypothesis is still tenable on observational grounds, it is not much help to the radio astronomer. The total kinetic energy of the collision of two massive galaxies would have to be converted into energy of relativistic electrons in order to explain the intense radio emission of Cygnus A and other intense sources. No method of accomplishing this is known.

Two other interesting suggestions have recently been made to explain the intense radio emission of Cygnus A and other peculiar galaxies. Geoffrey Burbidge of the Yerkes Observatory has suggested that a supernova explosion occurring in a region of high stellar density within a galaxy may trigger a chain reaction of other supernova explosions. Intense radiation from the first explosion falling on neighboring stars would cause nuclear reactions in the atmosphere of these stars. These nuclear reactions would in turn create disturbances that would travel inward to the center of the star and make it explode. The relativistic electrons needed for radio emission are produced either directly in the explosions or indirectly by collisions between cosmic ray protons produced in the explosions and the interstellar gas.

An entirely different proposal has been made by the Soviet physicist V. L. Ginzburg. Ginzburg suggests that intense radio emission may be associated with the formation of a galaxy. In this picture an original gas cloud—a proto-galaxy—contracts and breaks up into smaller gas clouds that subsequently form the stars of the galaxy. In the process gravitational energy is released that generates cosmic rays. Since the amount of interstellar gas would be large in the early life of a galaxy, before star formation had proceeded very far, collisions between cosmic rays and the gas atoms would be common. The collisions would provide the large number of relativistic electrons necessary for intense radio emission. If this hypothesis is correct, intense radio emission is a fairly short-



SPECTRAL DIFFERENCES are found between normal and peculiar radio galaxies. The normals (*black dots*) vary widely in spectral index and fall at the lower end of the intensity scale. Most peculiar radio galaxies (*color*) have a spectral index lying between -0.6 and -1.2 , and all emit substantially more radio energy than do the normal radio galaxies.

lived phenomenon that occurs rather early in the life of a galaxy.

Any theory for explaining the peculiar radio galaxies runs into two grave difficulties. What is the ultimate source of their enormous radio output? And how is that vast energy transported some tens of thousands of light-years from the galactic disk, allowing radio waves to originate far outside the visible limits of the galaxy? If the radiation is produced by synchrotron emission, the energy stored in the magnetic field and in relativistic electrons must be on the order of 10^{45} kilowatt hours (10^{60} ergs) for the most powerful radio sources. This is an appreciable fraction of the total nuclear energy of a galaxy. How are these electrons produced? If the electrons start off at low energy and are raised to relativistic energies by some acceleration process in the galaxy, the energy involved in the acceleration mechanism must be considerably greater than the final total energy of the relativistic electrons, because all known acceleration mechanisms are comparatively inefficient. If, on the other

hand, the electrons are produced as the result of collisions between cosmic ray protons and an interstellar gas, the total cosmic ray energy in the galaxy must also be at least one or two orders of magnitude greater than the relativistic electron energy. We are forced to conclude, therefore, that the mechanism underlying the intense radio emission from such objects as Cygnus A must be able to draw on a large fraction of the total energy of the object.

The two hypotheses mentioned above provide sources for the energy: nuclear reactions in one case and gravitational contraction in the other. Whether one or the other can also explain the observed distributions of radio emission and the various optical features, or whether additional theories will be required, remains to be seen. The unraveling of the processes that occur in radio galaxies has turned out to be an extraordinarily baffling problem. Its solution will undoubtedly lead to a deeper understanding of the history of galaxies and the universe.

The Author

D. S. HEESCHEN is an astronomer and currently acting director of the National Radio Astronomy Observatory in Green Bank, W. Va. Heeschén was born in Davenport, Iowa, in 1926, and he served with the Army Air Force in World War II before attending the University of Illinois, where he acquired a B.S. degree in 1949. He did graduate work in astronomy at Harvard University, was an Agassiz Fellow there from 1953 to 1954, taught for a year at Wesleyan University and received his Ph.D. from Harvard in 1955. The following year he spent as a lecturer at Harvard. He joined the staff of the National Radio Astronomy Observatory in 1956.

Bibliography

COSMIC RADIO WAVES. I. S. Shklovsky. Harvard University Press, 1960.

ON SYNCHROTRON RADIATION FROM MESSIER 87. Geoffrey Burbidge in *The Astrophysical Journal*, Vol. 124, No. 2, pages 416–429; September, 1956.

PARIS SYMPOSIUM ON RADIO ASTRONOMY, 1958. Edited by Ronald N. Bracewell. Stanford University Press, 1959.

POLARIZATION AND COMPOSITION OF THE CRAB NEBULA. J. H. Oort and Th. Walraven in *Bulletin of the Astronomical Institutes of the Netherlands*, Vol. 12, No. 462, pages 285–308; May 5, 1956.

THE RADIO EMISSION FROM NORMAL GALAXIES. R. Hanbury Brown and C. Hazard in *Monthly Notices of the Royal Astronomical Society*, Vol. 122, No. 6, pages 479–490; 1961.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

SUPERCONDUCTING MAGNETS

by J. E. Kunzler and Morris Tanenbaum

Magnets have now been made with superconducting coils through which current flows without resistance and heating. This indicates that it is possible to build very large magnets requiring very little power.

Soon after Heike Kamerlingh Onnes discovered superconductivity in 1911 he considered building a superconducting electromagnet. A series of experiments had convinced the Dutch physicist that at temperatures near absolute zero the electrical resistance of certain metals vanished completely, and that literally no energy was needed to maintain a flow of current through them, once started. An electromagnet made of such materials would require only enough energy to cool it; for the rest, the current would circle endlessly and the magnetic field would be perpetually available. Onnes' experiments failed because his materials lost their superconducting property in the presence of a magnetic field of moderate strength. In other words, the electromagnets would have destroyed their own superconductivity.

In subsequent years the incentive for building superconducting magnets steadily increased as magnets grew more powerful and the prospect of not having to maintain their huge currents became more appealing. Today there are perhaps a dozen conventional magnets in existence capable of generating steady magnetic fields of 100,000 gauss. (The field between the poles of a toy horseshoe magnet is a few hundred gauss.) Such magnets are expensive and require enormous amounts of power. For example, a modified Bitter solenoid magnet at the Bell Telephone Laboratories, which is used in various physical studies, requires 1.6 million watts of power to produce a field of about 100,000 gauss. This is about a quarter of the power consumed by the entire laboratory. Moreover, nearly 1,000 gallons per minute of cooling water are needed to remove the heat generated in the magnet winding.

In principle no energy is necessary to sustain a magnetic field once it is estab-

lished; therefore all the energy converted to heat is a result of the resistance of the magnet winding. Obviously if a superconductor could be used for the windings, and if it remained superconducting in the maximum field to which it was subjected, the energy-dissipation problems could be avoided.

The maximum magnetic field in which a superconductor remains superconducting is known as the critical field. Its value varies with the material and depends on temperature. The critical field is at a maximum at a temperature of absolute zero, and for "ideal" superconductors it decreases with rising temperature on a parabolic curve. It is zero at the critical temperature of the superconductor: the temperature above which the material cannot be superconducting [see top illustration on page 679].

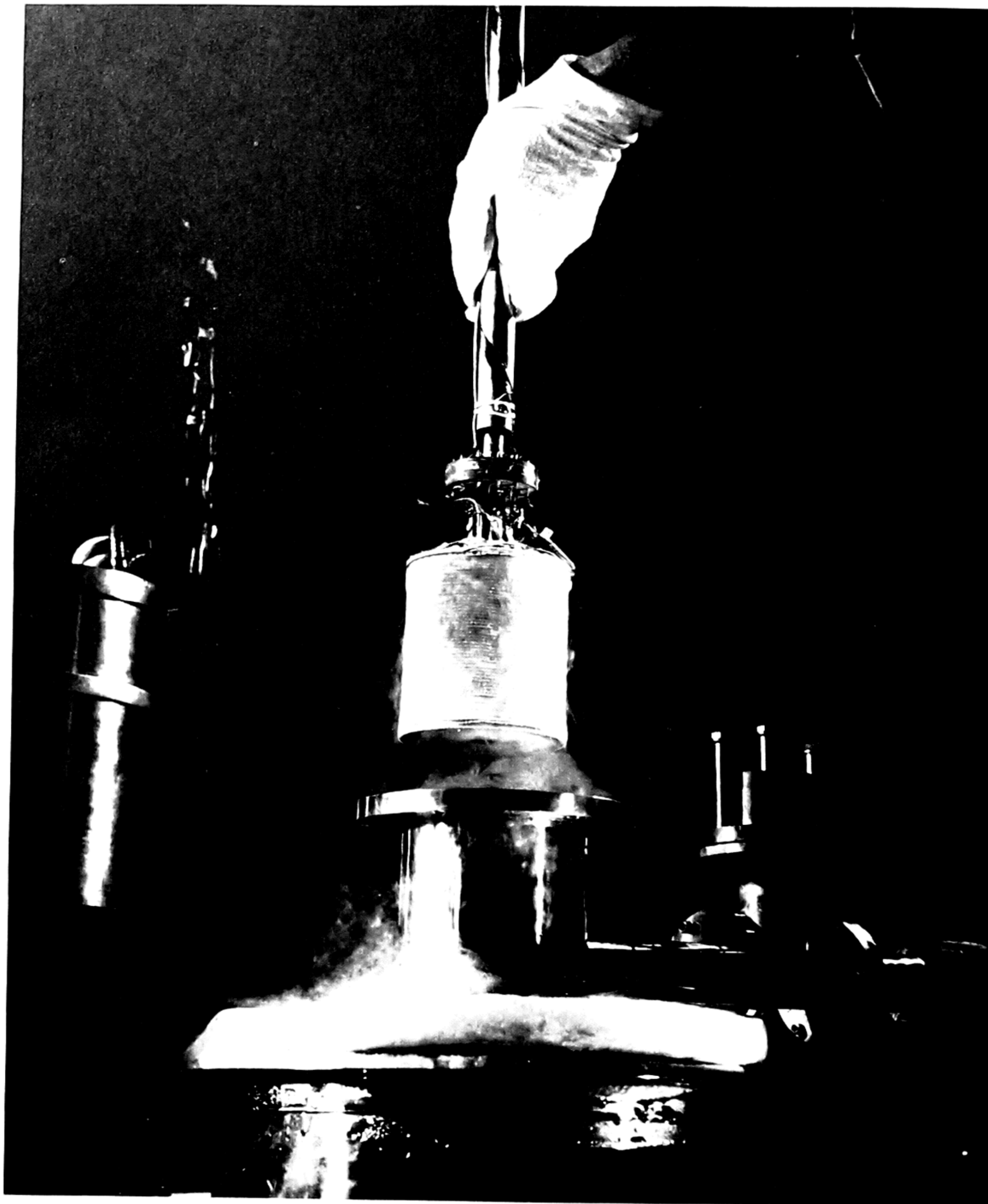
The superconductors available to Onnes and his immediate successors had low critical fields, of the order of a few hundred gauss or less. These materials acted as nearly ideal superconductors. It was also discovered that a sufficiently high current flowing through a wire could destroy superconductivity in the absence of an external magnetic field. For a time it appeared that the relation among critical field, critical current and temperature was reasonably well understood.

As alloys and mechanically hard superconductors came to be examined, however, it soon became apparent that these materials possess properties that are far from ideal. (The ideal superconductors became known as "soft" and the nonideal as "hard," chiefly on the basis of their mechanical properties. Although this correlation is a fair approximation, the names are misleading because there are many exceptions.) Among the hard superconductors traces of superconduc-

tivity persist in fields many times larger than those that had been expected from earlier studies.

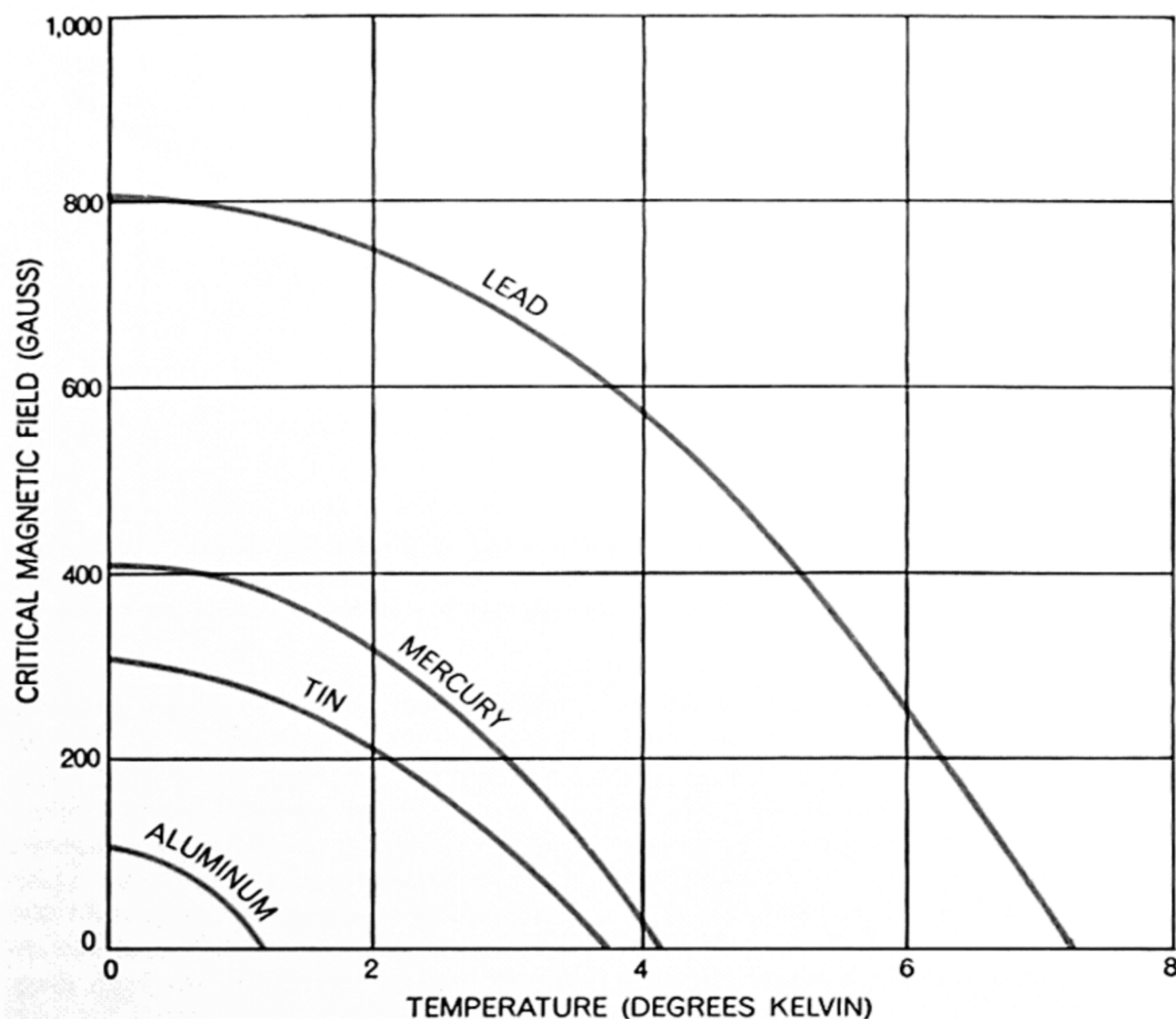
Superconducting magnets narrowly missed becoming a reality during the early 1930's. W. J. de Haas and J. Voogd at the University of Leiden found that wires of a hard superconducting alloy of lead and bismuth remained superconducting in fields in excess of 20,000 gauss. De Haas and Voogd accordingly supposed that it was possible to construct a superconducting magnet capable of generating a maximum field of about 20,000 gauss. Further experiments discouraged them from going ahead. Their colleague W. H. Keesom measured the maximum, or critical, current that the wire could carry and remain superconducting in a magnetic field. Extrapolating from the comparatively low fields in these studies, Keesom concluded that at high field values the critical current would be too small to be useful. (We now know that the critical current of a lead-bismuth alloy in fields of 10,000 to 15,000 gauss is sufficiently high to make this material practical for superconducting magnets capable of generating fields approaching 20,000 gauss.) As a result of this and related work it was widely concluded that superconducting magnets capable of fields of even a few thousand gauss were impractical. The miscalculation was unfortunate; the successful construction of superconducting magnets at that time would have had an impact on technology that would be hard to overestimate.

In 1934 Kurt A. G. Mendelssohn of the University of Oxford, making use of a theory of superconductivity developed by the physicists Fritz and Heinz London, was able to explain the existence of traces of superconductivity in fields of 10,000 to 20,000 gauss in terms of a

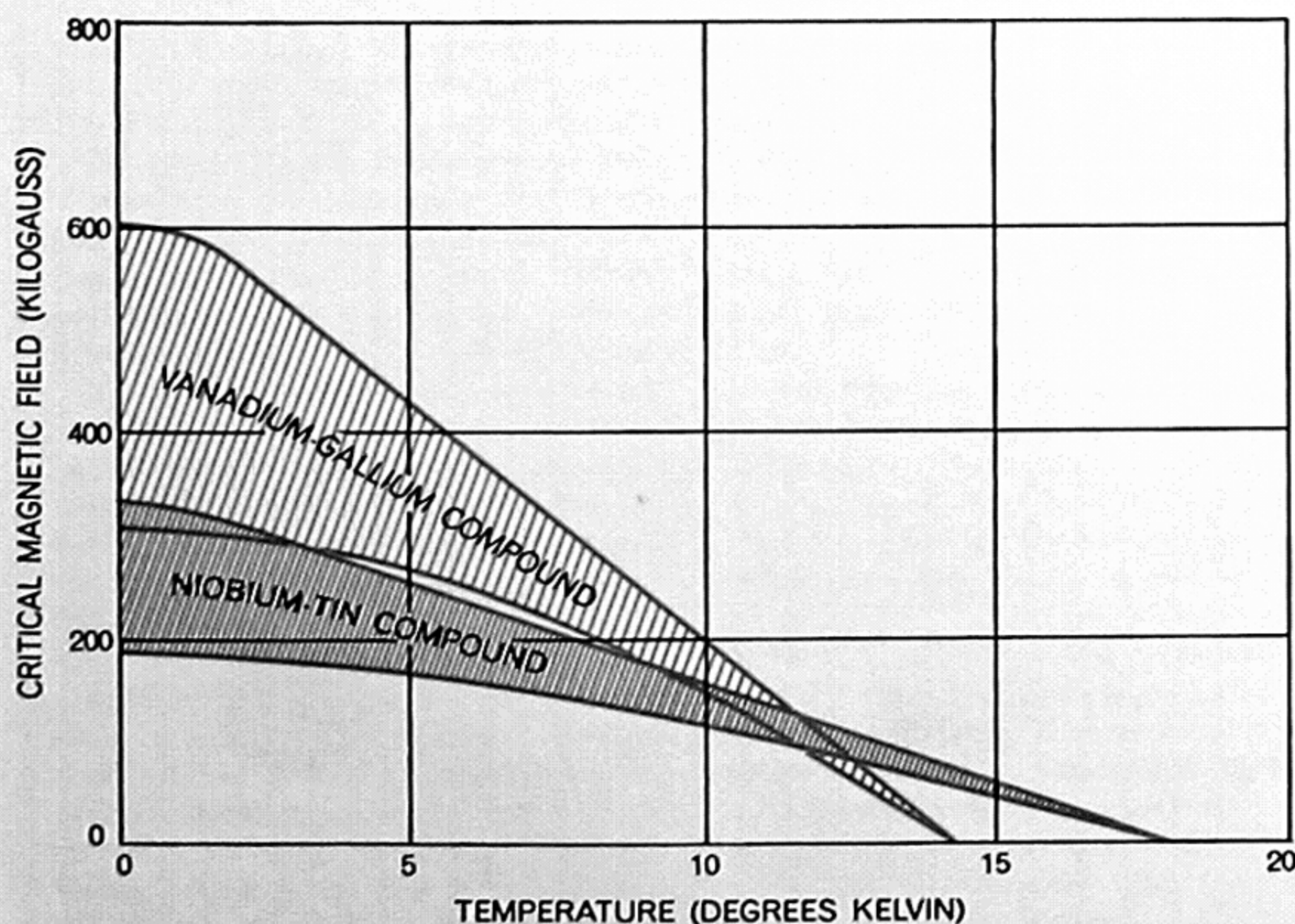


SUPERCONDUCTING MAGNET, the most powerful of several recently built at the Bell Telephone Laboratories, is shown being removed from a vessel containing liquid helium. When the magnet

is running, the vessel is closed and the deeply chilled niobium-tin coils of the device allow electric current to flow without resistance, producing a field of 70 kilogauss (70,000 gauss) inside the magnet.



CRITICAL MAGNETIC FIELD is the maximum field in which a superconductor remains superconducting at a given temperature. The critical temperature is the maximum temperature for superconductivity at zero field. The interrelations are plotted here for four "ideal," or "soft," superconductors. They cannot be used for making high-field superconducting magnets because their superconductivity is killed by low current and low magnetic fields.

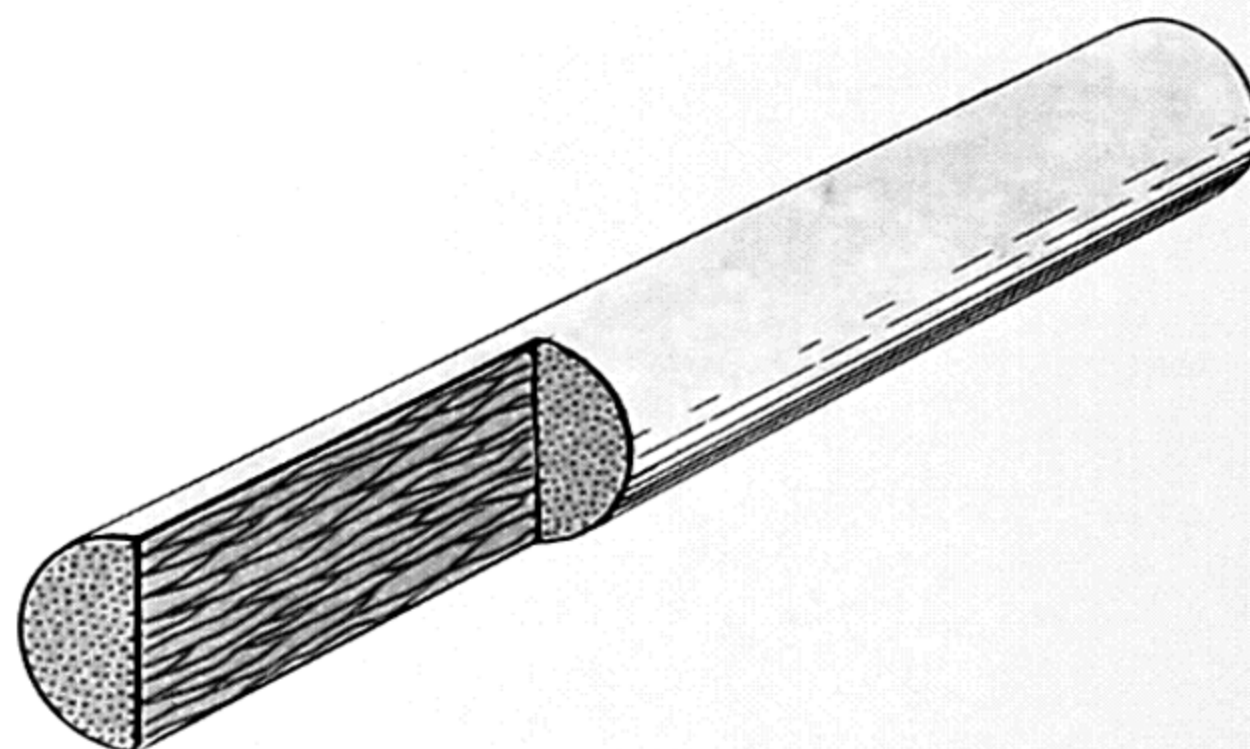
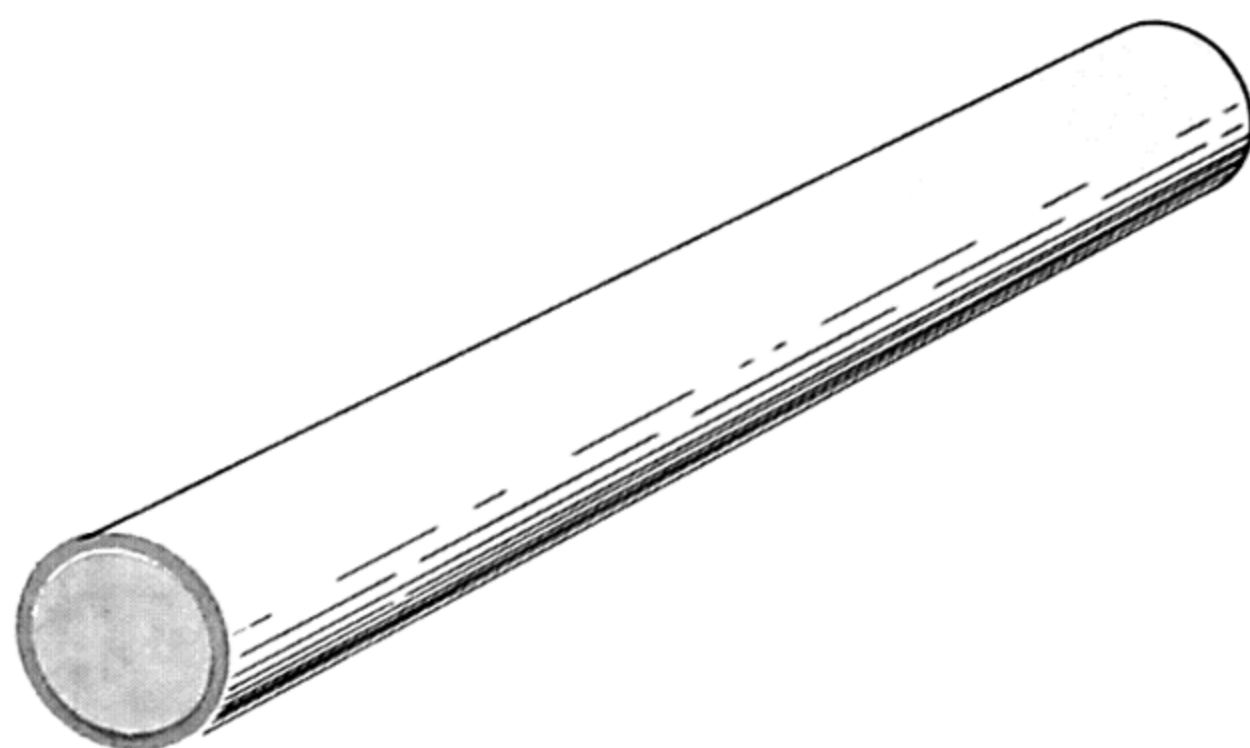


VERY HIGH CRITICAL FIELDS can be sustained by certain "nonideal," or "hard," compounds, such as those consisting of vanadium and gallium (V_3Ga) or niobium and tin (Nb_3Sn). Their critical fields can be estimated only between broad limits (shaded) because there are no magnets capable of testing their properties much above 100 kilogauss. These compounds have the virtue of remaining superconducting while carrying high currents.

spongelike or a three-dimensional filamentary structure. In an ideal superconductor the current is carried only within a thin surface layer, known as the penetration depth, with a thickness of the order of a few hundred angstrom units (ten-millionths of a millimeter). The theory predicted that for films of material having a thickness less than a certain critical amount, the critical field increased as the layers got thinner. In other words, a very thin film has a higher critical field than a piece of bulk material. Similar conclusions apply to thin wires. Therefore a three-dimensional network consisting of thin filaments, or thin films, with a thickness much less than the critical amount could be expected to remain superconducting in magnetic fields several times stronger than the critical fields for the corresponding bulk materials. Mendelssohn assumed that only a small fraction of the bulk of the hard superconductor is contained in the filaments. The magnetic field can penetrate both the nonsuperconducting part of the material and the filaments because they are so thin.

Even before Mendelssohn's work it had been known that the superconducting properties of some materials could be enhanced by deforming the superconductor mechanically. In 1955 G. B. Yntema, then at the University of Illinois, used this stratagem to make a small electromagnet with an iron core and a superconducting coil of niobium wire. Subsequently J. K. Hulm and his associates at the Westinghouse Research Laboratories successfully operated air-core niobium magnets. In 1960 Stanley A. Autler of the Lincoln Laboratory of the Massachusetts Institute of Technology described the first practical applications of a superconducting niobium magnet. Autler's magnet generated a field of 4,300 gauss when it was cooled to the boiling point of helium (4.2 degrees centigrade above absolute zero) and was used in a solid-state microwave maser.

At the Bell Telephone Laboratories in 1960 B. T. Matthias suggested using an alloy of molybdenum and rhenium to make superconducting magnets. (Hulm had reported alloys of these metals to be superconducting five years earlier.) One of us [Kunzler], working with Ernest Buehler, F. S. L. Hsu and Charles Wahl, examined the critical current as related to the magnetic field in the alloys. The Bell Laboratories group was able to show that it should be possible to construct a solenoid, using molybdenum-rhenium wire, that would be capable of



SOFT AND HARD SUPERCONDUCTORS differ in the manner in which they carry electric current. In a soft superconductor (*left*) current is carried only in a thin surface layer. In hard

superconductor (*right*), such as niobium-zirconium, the current seems to be carried by filaments inside the material. Number of filaments can be increased by physically working the material.

producing a superconducting field of approximately 15,000 gauss. They then constructed such a solenoid and found that its operating characteristics were consistent with the measurements on small samples of the wire.

In 1954 Matthias, T. H. Geballe, Seymour Geller and Ernest Corenzwit reported the synthesis of niobium-tin

(Nb_3Sn), a compound that becomes superconducting at 18 degrees centigrade above absolute zero. Today niobium-tin remains the material with the highest known temperature of transition from ordinary conductivity to superconductivity. In February, 1961, Kunzler, Buehler, Hsu and Jack H. Wernick pointed out that this compound satisfied

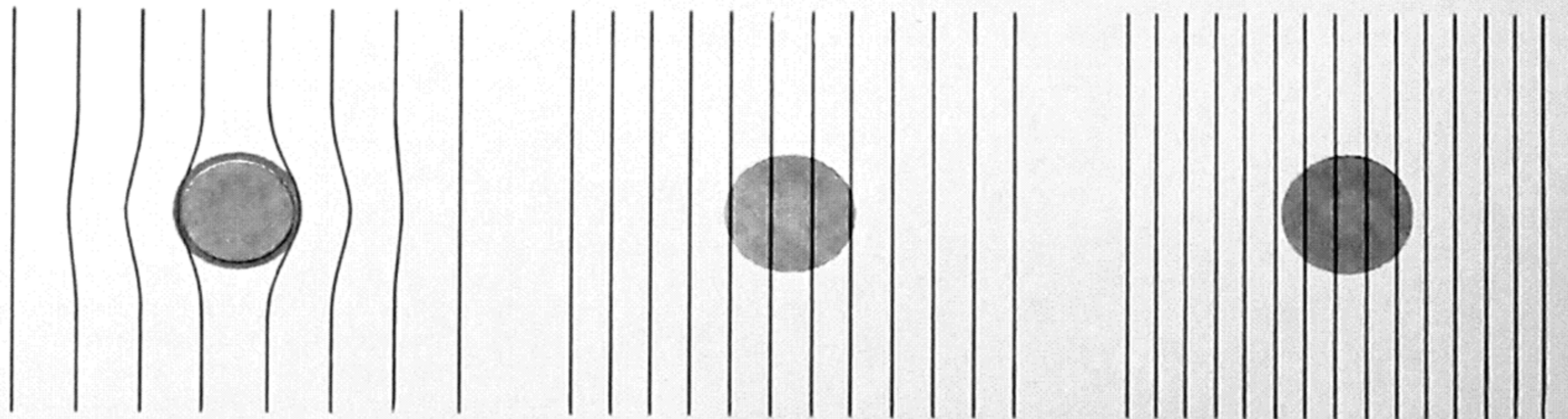
spectacularly well the three essential requirements of a material suitable for construction of superconducting magnets: (1) The material must remain superconducting in a high magnetic field; niobium-tin was found to be superconducting in a magnetic field of 88,000 gauss. (2) The material must sustain a high current density in the high mag-

LOW FIELD
(UNDER .1 KILOGAUSS)

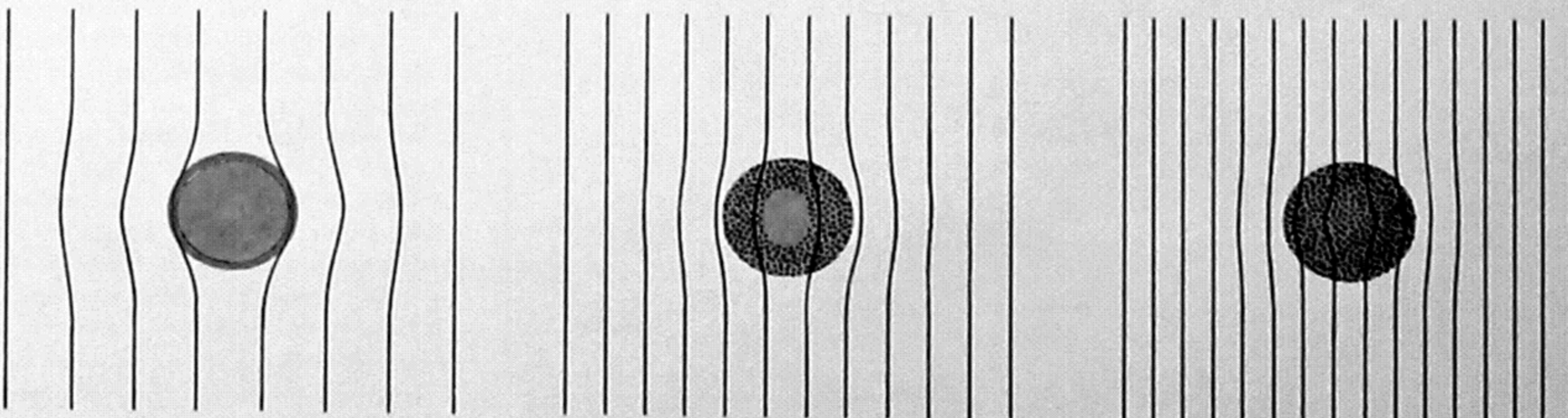
MODERATE FIELD
(1 TO 10 KILOGAUSS)

HIGH FIELD
(OVER 10 KILOGAUSS)

a

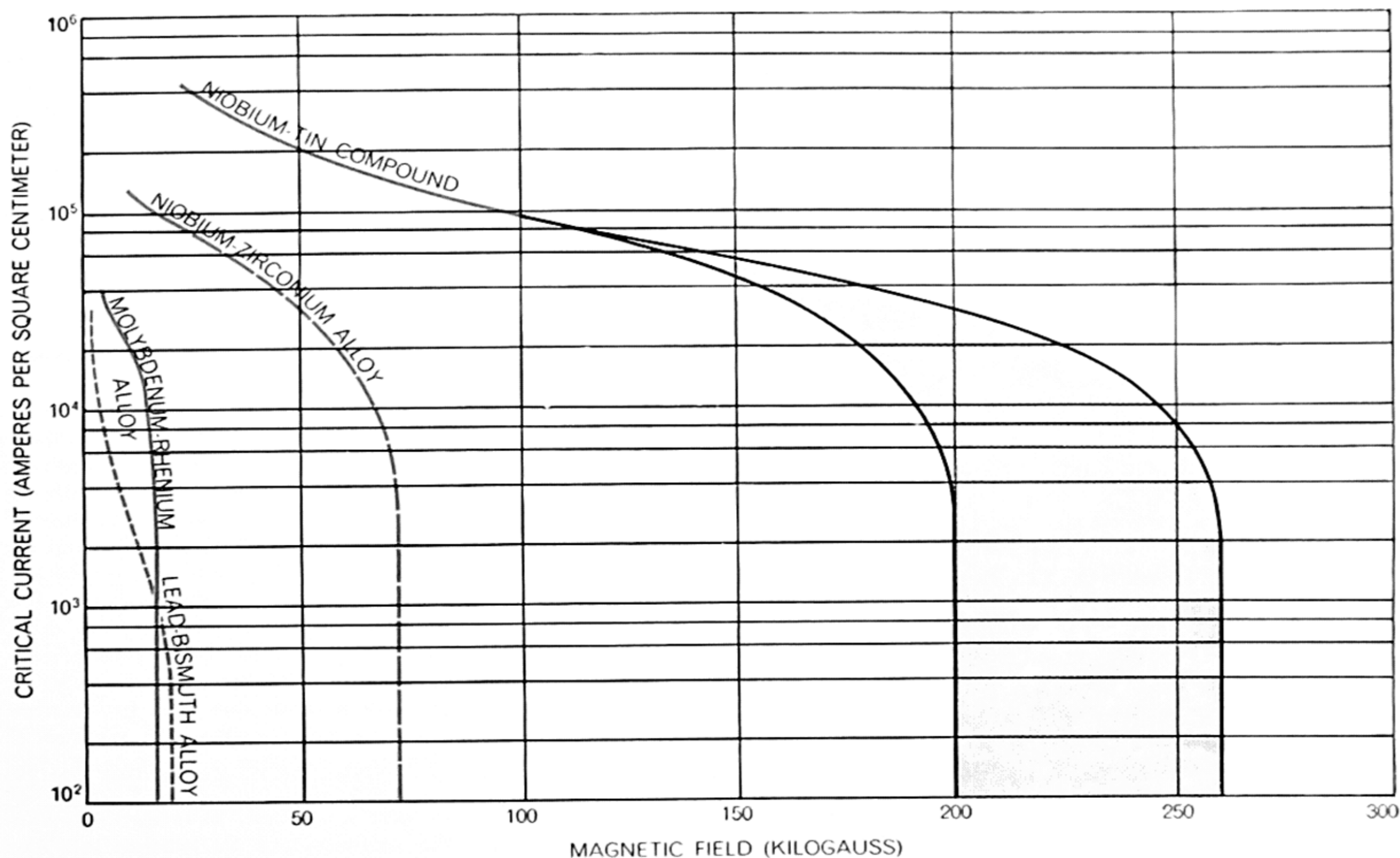


b



EFFECT OF MAGNETIC FIELDS on a typical soft superconductor (*a*) is different from that on a hard superconductor (*b*). When the field is low, under .1 kilogauss, both conduct current in a thin surface layer (*color*), and the field is excluded from the interior of the material. A moderate field, one to 10 kilogauss,

penetrates a soft superconductor, thereby destroying its superconductivity. When a moderate field penetrates a hard material, however, superconducting flow appears to be transferred to thin filaments. In certain hard superconductors filamentary flow persists even in high fields of 10 to 100 kilogauss and beyond.



CRITICAL CURRENT DENSITY is the maximum current a superconductor can carry and still remain superconducting. When the superconductor is placed in an increasing magnetic field, the critical current density falls slowly at first and then precipitously, as shown by these four curves. A superconducting magnet requires a

material that maintains a high critical current density in the presence of high magnetic fields. Plotted values are for materials cooled to the boiling point of liquid helium, 4.2 degrees Kelvin (*see top illustration on following page*). Magnets do not yet exist for measuring critical current density much beyond 100 kilogauss.

netic field; niobium-tin wire was shown to sustain a current density in excess of 100,000 amperes per square centimeter at 88,000 gauss. (3) The material, even if refractory, must be capable of being fabricated into a magnet; niobium-tin is extremely brittle, but it can be formed into solenoids by a special mode of manufacture that will be described later.

The observation that niobium-tin had exceptional properties was a surprise, but it came at a time when technology was ready to take full advantage of the fact. First, there was a need for superconducting magnets both as research tools and for devices such as the solid-state maser. The desire for a superconducting magnet to be used with a solid-state maser had stimulated Autler's work. It was a similar motive that had initially interested the Bell Laboratories workers in superconducting magnets. Second, a large amount of information on superconducting materials had accumulated, thanks to the work of Matthias and Hulm; one or the other of them discovered superconductivity in all the materials of greatest interest for magnet construction. Third, the tremendous re-

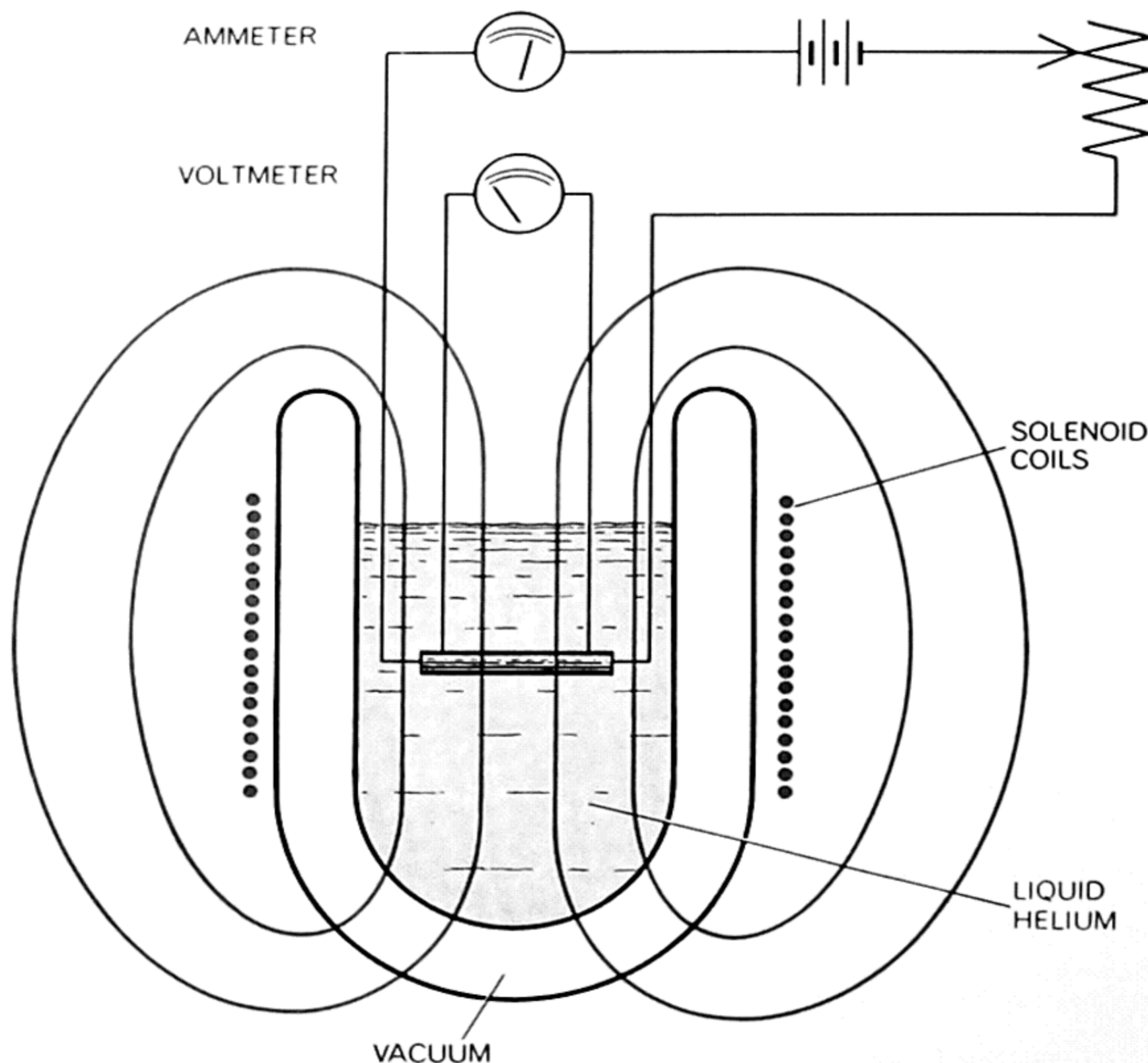
cent advances in low-temperature technology make possible arrangements that would have been out of the question a few years ago. At that time liquid helium was a curiosity available only in a few laboratories. Today it can be bought commercially and is on hand in nearly every major laboratory.

In investigating the properties of superconductors and their possible application to magnets it is not practical actually to build a magnet of each of the possible materials. Most of the potentially interesting materials (such as niobium-tin) are not readily available in suitable form. In fact, they are usually quite difficult to prepare, and the development of fabrication procedures is warranted only if a material shows promise. The best test is to determine how the critical current density decreases as the applied magnetic field is increased. The sample is cooled to liquid helium temperatures and a transverse magnetic field is applied. Then current is passed through the material in gradually increasing quantity until a voltage suddenly appears across the sample. This signals the

appearance of resistance and therefore the end of a continuous superconducting pathway.

Typical curves for four materials are shown above. All have the same general shape: a "knee" followed by a rapid decrease in critical current with increasing magnetic field. The maximum practical field for which a material can be used in magnet construction corresponds approximately to the field at the knee of the curve. This value has not been directly measured for many of the new compounds because sufficiently strong steady-state fields are not available. Niobium-tin has been tested in such fields only up to about 100,000 gauss, and the knee lies somewhere beyond this value. Some experiments indirectly suggest a value in the neighborhood of 200,000 gauss. Therefore niobium-tin magnets should be able to generate fields of about 200,000 gauss.

Niobium-zirconium, which forms ductile alloys, was first reported to be superconducting by Matthias in 1953. At the April, 1961, meeting of the American Physical Society, a group of Bell Laboratories workers drew attention to its high-



APPARATUS FOR EVALUATING SUPERCONDUCTORS requires the use of a variable magnetic field. Measurements are made on a sample held at right angles to field and cooled to temperature of liquid helium. As long as the sample superconducts it offers no resistance and the voltmeter reads zero. Raising either the current flow or the magnetic field, or both, ultimately destroys superconductivity, which is reflected by a voltage reading.

field superconducting properties and its usefulness for magnets. The position of the knee for niobium-zirconium alloys depends on the composition and increases with increasing zirconium content, up to a proportion of zirconium of about 75 per cent. The optimum critical current density, however, is found in a compound containing between 25 and 35 per cent zirconium. The maximum field for which these alloys may prove useful for magnets is between 80,000 and 100,000 gauss.

The critical current of any material is highly sensitive to its state of mechanical deformation. The greater the deformation that is brought about by cold-working, the greater, as a rule, the critical current-carrying capacity. On the other hand, the maximum field at which superconductivity persists near zero current density is not affected by the mechanical state of the materials. This and related evidence has provided a clue as to the origin of the filaments responsible for the existence of high critical fields. The filaments are thought to be associated with dislocations in the crystalline structure of the material: defects in the lattice

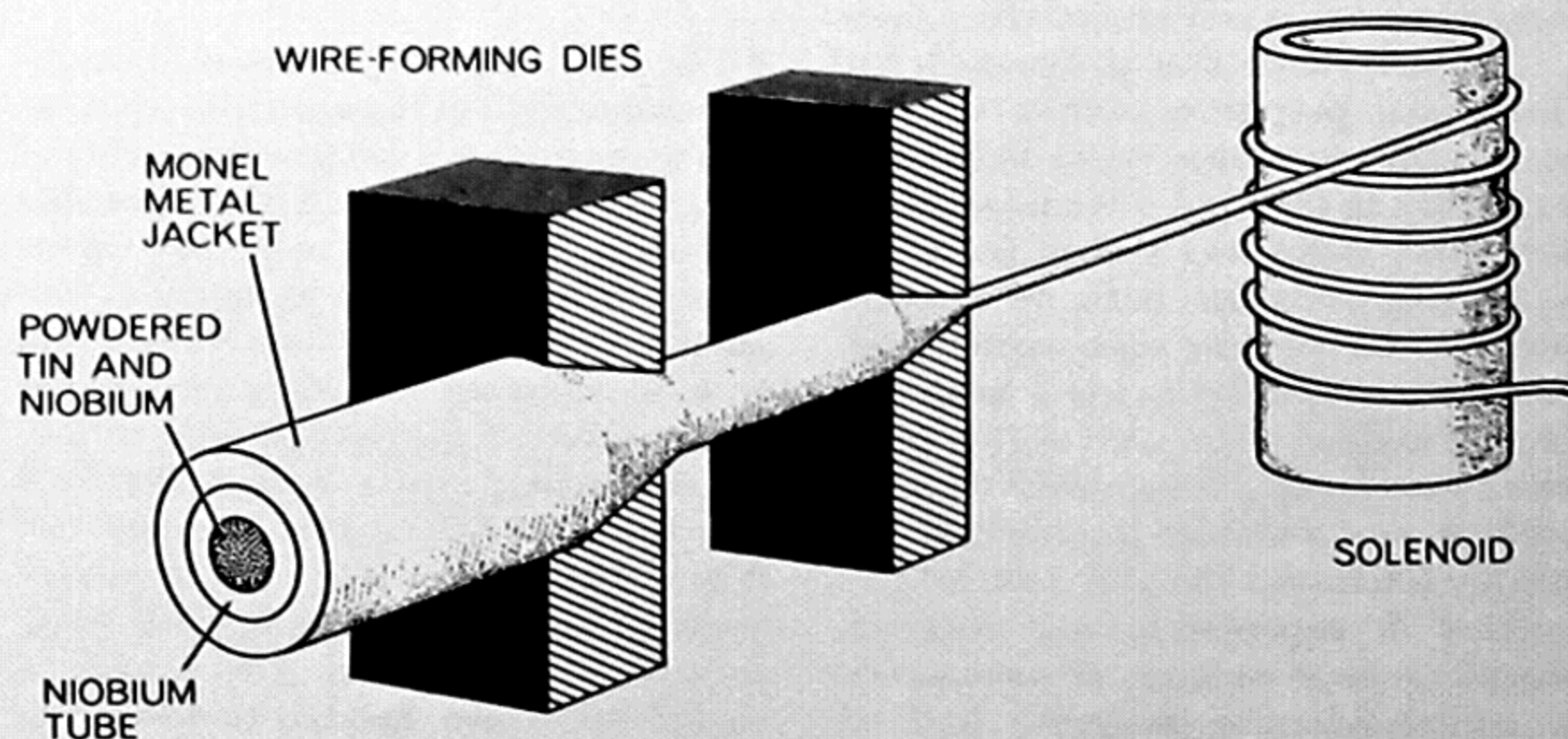
of a crystal that increase in number when the crystal is deformed.

The structure of individual dislocations probably does not change with increasing deformation; their effective diameter does not increase, only their number. It is the diameter, however, that

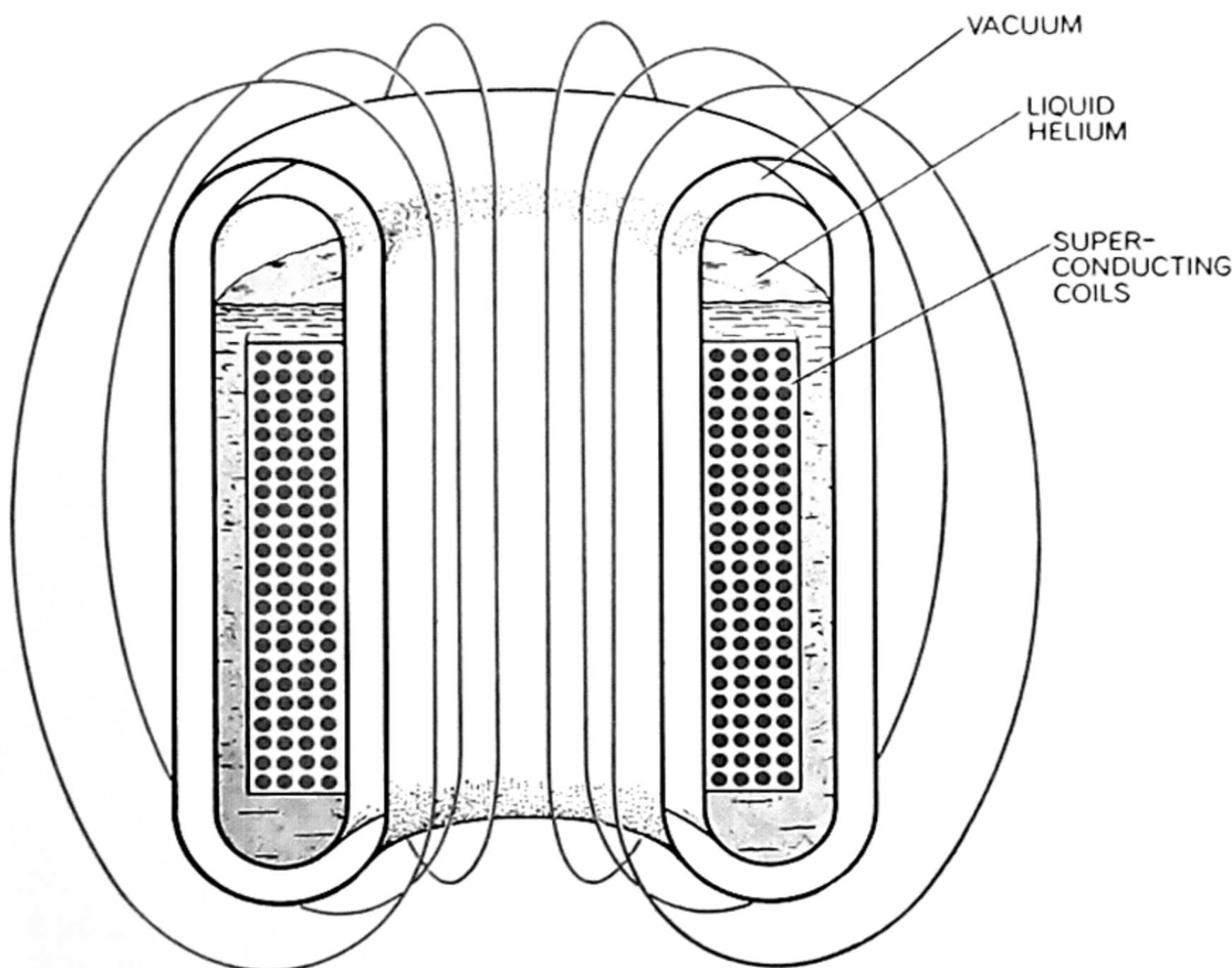
presumably determines the thickness of the filament and therefore the maximum field in which superconductivity can exist. Thus the increase in critical current density with mechanical deformation and the lack of increase in the limiting magnetic field are consistent with the dislocation interpretation. This model has the further appealing feature that dislocations do not terminate within one of the many single-crystal grains that make up a metal but can form a continuous and connected network between grain boundaries.

We have mentioned that special techniques are required to fabricate niobium-tin into magnets. A niobium tube is filled with an intimate mixture of finely divided niobium powder and tin powder. Then the ends of the tube are sealed with niobium plugs and the tube is drawn into "wire." In a typical wire the outside diameter of the niobium jacket is about .015 inch; the diameter of the core is about .006 inch. In this state the core still consists of a mixture of elemental niobium and elemental tin; the wires are ductile and can be wound to form a solenoid. In order to convert the niobium and tin mixture into the compound Nb_3Sn , the coil is heated in a furnace at 1,000 degrees centigrade. Once the compound core has formed, the wire is brittle and cannot be bent without damaging it. But it is already in the required shape. Several 10,000-foot lengths of shaped niobium-tin wire have been prepared by Karl M. Olsen, Robert F. Jack and Edward O. Fuchs of the Bell Laboratories in collaboration with the Superior Tube Company and the Wilbur B. Driver Company.

Because the wire must be heated to



NIOBIUM-TIN WIRE for use in superconducting magnets is made by drawing down a tube containing niobium and tin in the form of intimately mixed powders. The resulting wire is wound into solenoid coils. The coils are then heated to about 1,000 degrees centigrade, whereupon niobium and tin unite chemically to form Nb_3Sn , a highly brittle compound.



WARM-CORE SUPERCONDUCTING MAGNET could be built by isolating the low-temperature components. Core, containing the intense field, could then be any temperature.

so high a temperature in the process of fabrication, the problem of providing it with an insulating coating is not easy. It was found that covering the original cylinder with Monel metal not only insulates the drawn wire but also aids the drawing. Geballe first pointed out that a layer of metal would adequately insulate a superconductor. The resistance of the layer is infinite compared with that of the superconductor.

Thus when the insulated wire is formed into a solenoid coil and the flow of current through it reaches a steady state, all the current passes through the superconducting core and none through the Monel metal coating. When the current is first turned on, however, its build-up in the solenoid is impeded not primarily by electrical resistance but by inductance, or "electrical inertia." During the time that the current is increasing, the inductance of the superconducting core will force a large part of the flow into the Monel metal insulation. If the adjacent turns are in direct contact, the solenoid will be effectively short-circuited. To avoid this the layers of the Monel-metal-clad wire are separated by layers of quartz-fiber cloth.

At the time of publication of this article the general properties of niobium-tin and niobium-zirconium will have been known for only about a year. Already superconducting magnets of

both materials have been constructed and operated in fields of approximately 70,000 gauss. At least three firms are offering for sale superconducting magnets capable of producing fields up to 50,000 gauss. A 70,000-gauss niobium-tin experimental magnet has been tested at the Bell Laboratories, and a 70,000-gauss niobium-zirconium magnet at the Westinghouse Research Laboratories. Atomics International (a division of North American Aviation, Inc.) and the Lincoln Laboratory are among the other organizations that have reported magnets capable of producing fields above 50,000 gauss. This is remarkable progress, yet it probably represents only a fraction of what can be accomplished in another year or so.

In the experimental magnets built so far the diameter of the working space inside the solenoid is usually a fraction of an inch. Such small dimensions do not reflect any fundamental limitation. These are early prototype magnets built primarily to test materials and design.

Although niobium-tin, with a potential field of about 200,000 gauss, will probably be used in the first superconducting magnets operating near 100,000 gauss, other materials offer the possibility of reaching even higher fields. One of the most intriguing of these is vanadium-gallium (V_3Ga). Like so many of the others, this compound was first made

and found to be superconducting by Matthias. Wernick, F. J. Morin and their associates at the Bell Laboratories recently obtained experimental evidence indicating that the critical field of vanadium-gallium is in excess of 400,000 gauss. Since the study of high-field superconductors is still in its infancy, it seems reasonable to expect the discovery of materials with still better properties.

Although the vanadium-gallium data suggest that it is theoretically possible to construct superconducting magnets capable of generating fields of several hundred thousand gauss, there are many problems to be solved before such magnets become a reality. Not the least of these is to contain the forces that are generated. For example, a field of 300,000 gauss produces an outward pressure of about 50,000 pounds per square inch.

Since all known superconductors exhibit superconductivity only in the neighborhood of the temperature of liquid helium, it might seem at first that any device used in conjunction with a superconducting magnet would also have to be maintained at very low temperatures. This is not the case, however. It is not difficult to arrange a solenoid cooled by liquid helium so that its core is maintained at room temperature or at an even higher temperature.

Superconducting magnets appear to be potential candidates for almost every application involving large magnetic fields. Among the obvious possibilities are the furnishing of fields for solid-state physics research and communication devices, for particle accelerators such as cyclotrons, and for the bending of the paths of charged nuclear particles in detectors such as cloud chambers and bubble chambers. Superconducting magnets are particularly intriguing in the field of power generation, both for magneto-hydrodynamic devices and for controlled nuclear fusion.

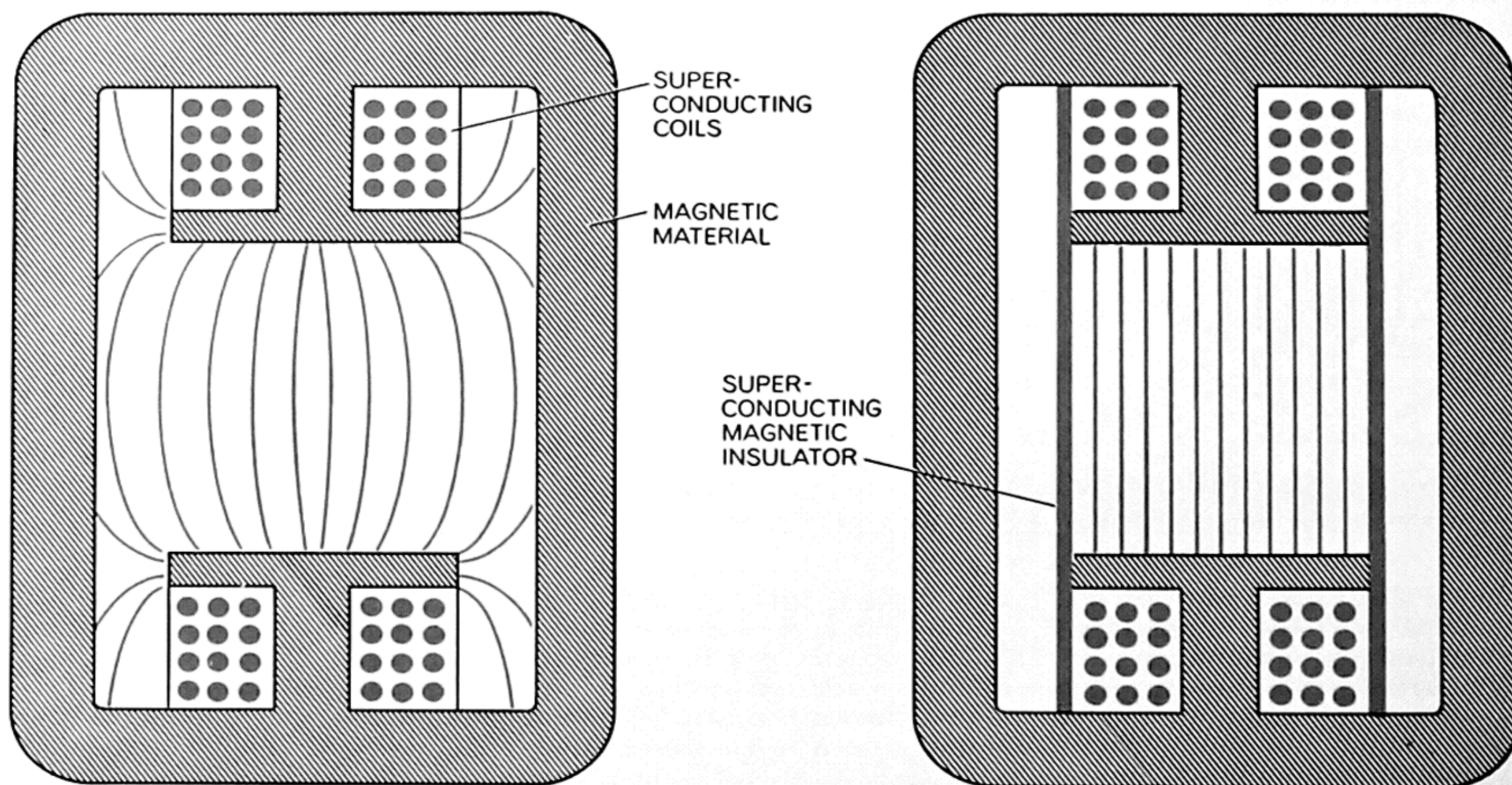
This last application is one of the most interesting and potentially the most important. There are many problems that must be solved before fusion power becomes a practical reality. One is the confinement of hot ionized gases, or plasmas, in some sort of container. Because the plasmas will be at temperatures in the range of 100 million degrees centigrade, no material substance can be used to contain them. They can, however, be confined by the force of a magnetic field. A longitudinal field can form a "magnetic bottle" and keep the hot ionized gases away from the walls of the physical container [see lower illustration on following page]. In order to

keep the plasma from escaping from the ends, a higher magnetic field is provided, forming constrictions in these regions. Until recently it was planned to use an ordinary good conductor, such as copper, for the magnet to provide the necessary fields. But such magnets would be massive—possibly several feet in diameter and tens of feet long—and they would have to create magnetic fields

of several tens of thousands of gauss.

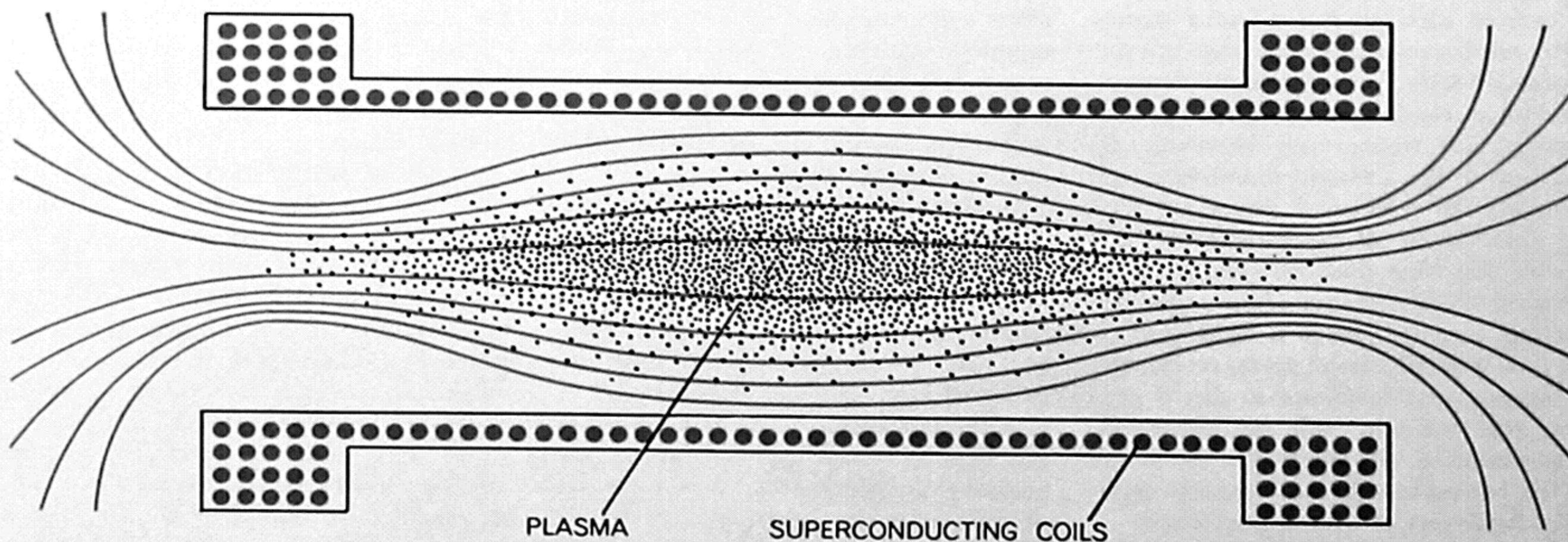
Estimates have shown that the amount of energy required to operate magnets of this size will exceed the output of the thermonuclear power generator unless it is made extremely large. Even then a tremendous amount of power will be required initially to start the generator, and the source of this starting power presents a problem. Current thinking in-

volves the use of superconductors to provide the magnetic field. The power required to sustain the magnetic field will be negligible. Power will be required, it is true, to refrigerate the windings to liquid-helium temperatures. Although this power will be appreciable, it will be trivial compared with the power required for the conventional type of magnet.



SHAPED-FIELD SUPERCONDUCTING MAGNET was designed by P. P. Cioffi of the Bell Telephone Laboratories for use with a solid-state maser, a device for amplifying radio signals. The magnet, shown in cross section at right, combines superconducting coils

with superconducting magnetic insulators. The latter shape the magnetic field by blocking its passage. If the insulators were not present, the field would take the form shown at left. The 10-pound superconducting magnet replaces a conventional one of 700 pounds.



"MAGNETIC BOTTLE," of the type employed in studying controlled thermonuclear reactions, could be built with superconducting coils. The purpose of the bottle is to confine ionized gas, or plasma, which has been heated to millions of degrees centigrade.

The ionized particles are trapped by the magnetic lines of force and so are kept from touching the walls of the container. When the particles encounter the stronger magnetic fields at the ends of the container, they are "reflected" back toward the middle.

The Authors

J. E. KUNZLER and MORRIS TANENBAUM are respectively head of the Metal Physics Research Department and assistant director of the Metallurgical Research Laboratories at the Bell Telephone Laboratories. Kunzler received a B.S. in chemical engineering from the University of Utah in 1945 and a Ph.D. in physical chemistry from the University of California in 1950. He was a research associate at California until 1952, when he went to the Bell Laboratories. There Kunzler established a low-temperature laboratory and began specializing in the precise measurement of the electrical, thermal and mechanical properties of solids at low temperatures. This led him to the study of superconductivity and superconducting magnets. Tanenbaum obtained an A.B. at Johns Hopkins University in 1949 and a Ph.D. in chemistry at Princeton University three years later. His first years at the Bell Laboratories, which he joined in 1952, were devoted to studies of the chemistry and physics of semiconduc-

tors. Tanenbaum's later interest in the structural and electrical properties of solids led to his association with Kunzler in work on the properties of superconductors in strong magnetic fields.

Bibliography

PROGRESS IN LOW TEMPERATURE PHYSICS, VOL. II. Edited by C. J.orter. Interscience Publishers, Inc., 1957.

SUPERCONDUCTING MATERIALS AND HIGH MAGNETIC FIELDS. J. E. Kunzler in *Journal of Applied Physics*, supplement to Vol. 33, No. 3, pages 1042-1048; March, 1962.

SUPERCONDUCTIVITY. D. Shoenberg. Cambridge University Press, 1952.

SUPERCONDUCTIVITY. B. T. Matthias in *Scientific American*, Vol. 197, No. 5, pages 92-103; November, 1957.

SUPERCONDUCTIVITY IN HIGH MAGNETIC FIELDS AT HIGH CURRENT DENSITIES. J. E. Kunzler in *Reviews of Modern Physics*, Vol. 33, No. 4, pages 501-509; October, 1961.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

INCLUSION COMPOUNDS

by John F. Brown, Jr.

They are substances in which a molecular structure encloses atoms or molecules of another kind. They are finding uses in industry and play a role in the chemistry of the living cell.

For half a century chemists speculated that certain molecular structures might enclose other structures of suitable size and geometry. It was not until the late 1940's, however, that this form of molecular architecture was actually shown to exist. Then, as so often happens when the time is ripe, the discovery was made almost simultaneously by several different research groups. The combination is known as an inclusion compound. By 1952 examples of all six basic types of inclusion compound had been found, and since then their characteristics have been intensively studied.

Inclusion compounds are chemical combinations in which one component fits into a cavity in the other. When the size and shape of the cavity in the "host" molecule just match the form of the "guest," a combination of appreciable strength can occur. No ordinary chemical bonds between the atoms of the host and guest are needed. In well-fitted inclusion structures, in which many atoms of the host molecule are close to those of the guest, a sizable total binding force is provided by the interactions of the outer electrons in the shells of the host and guest atoms. (Such binding forces are known as dispersion forces.) The structural arrangements that allow this snug "hand in glove" fitting of molecular shapes differ from one family of inclusion compounds to another. We shall discuss the various families in turn.

Molecular Traps

It has been known for a century that quinol, or hydroquinone, widely used as a photographic developer, produces curious "complexes" when crystallized out of certain solutions. These complexes contain one molecule of liquid

or gas for each three of quinol. They are perfectly stable in the dry state at room temperature, and have no smell of the occluded gas, even when it is the odoriferous hydrogen sulfide or sulfur dioxide. When the complex is melted or dissolved in water, the occluded gas is immediately evolved, showing that it is not held by stable chemical bonds.

Between 1947 and 1950 H. M. Powell of the University of Oxford reported the structures of various quinol complexes. He found that the host molecules are linked by hydrogen bonds between oxygen atoms (that is, by bonds in which a hydrogen atom acts as a bridge between two oxygen atoms) to form a pair of interlocking three-dimensional networks. The networks do not, however, completely fill the available space. The cavities remaining are roughly spherical and about four angstrom units in diameter. (An angstrom is one hundred-millionth of a centimeter.) Each cavity is bounded by two circles of six hydrogen-bonded hydroxyl groups and by the benzene rings of six quinol molecules. The guest molecules lie trapped in these cavities [see top illustration on page 688]. Powell suggested the apt term "clathrate" (from the Latin word *clathratus*, which means enclosed by the bars of a grating) to describe this type of combination.

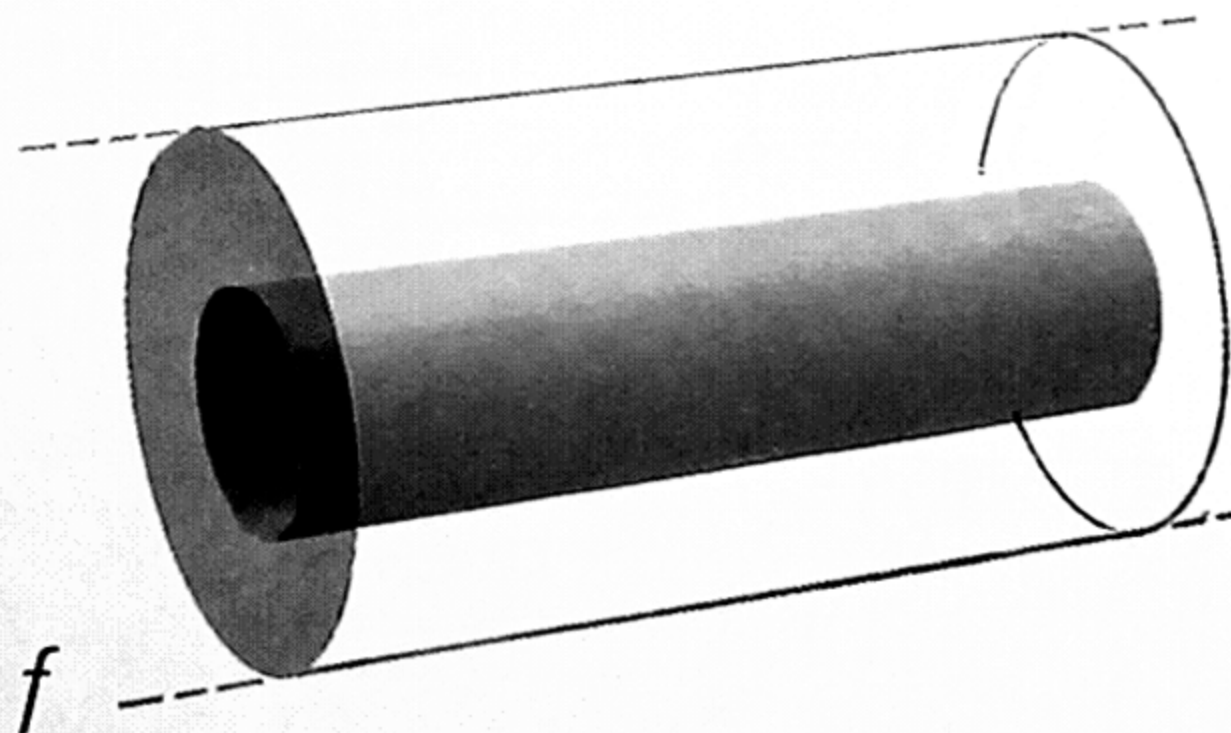
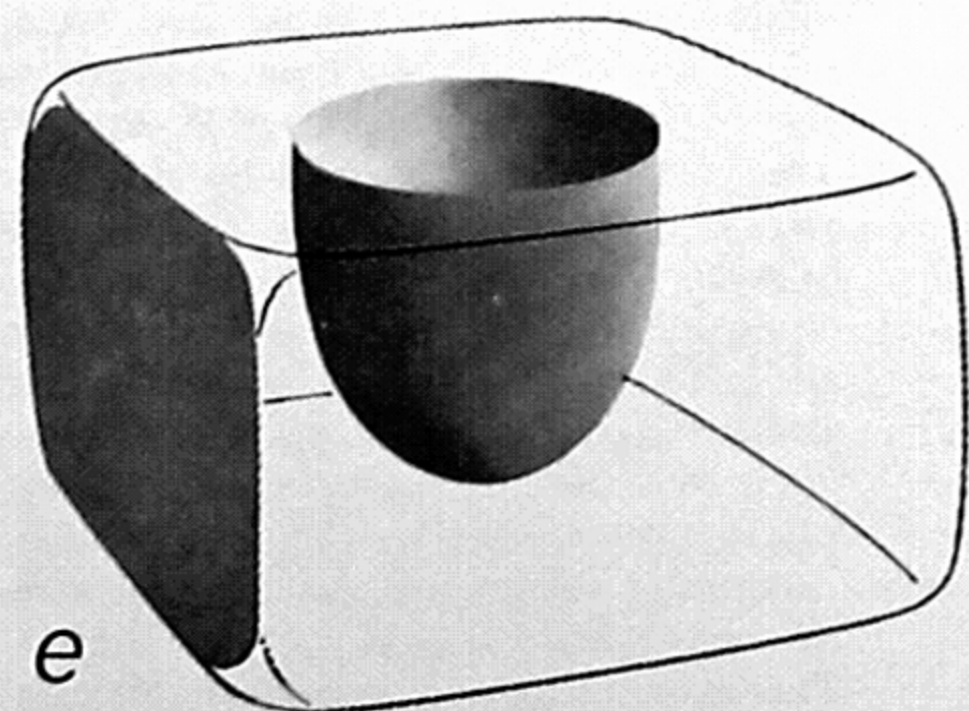
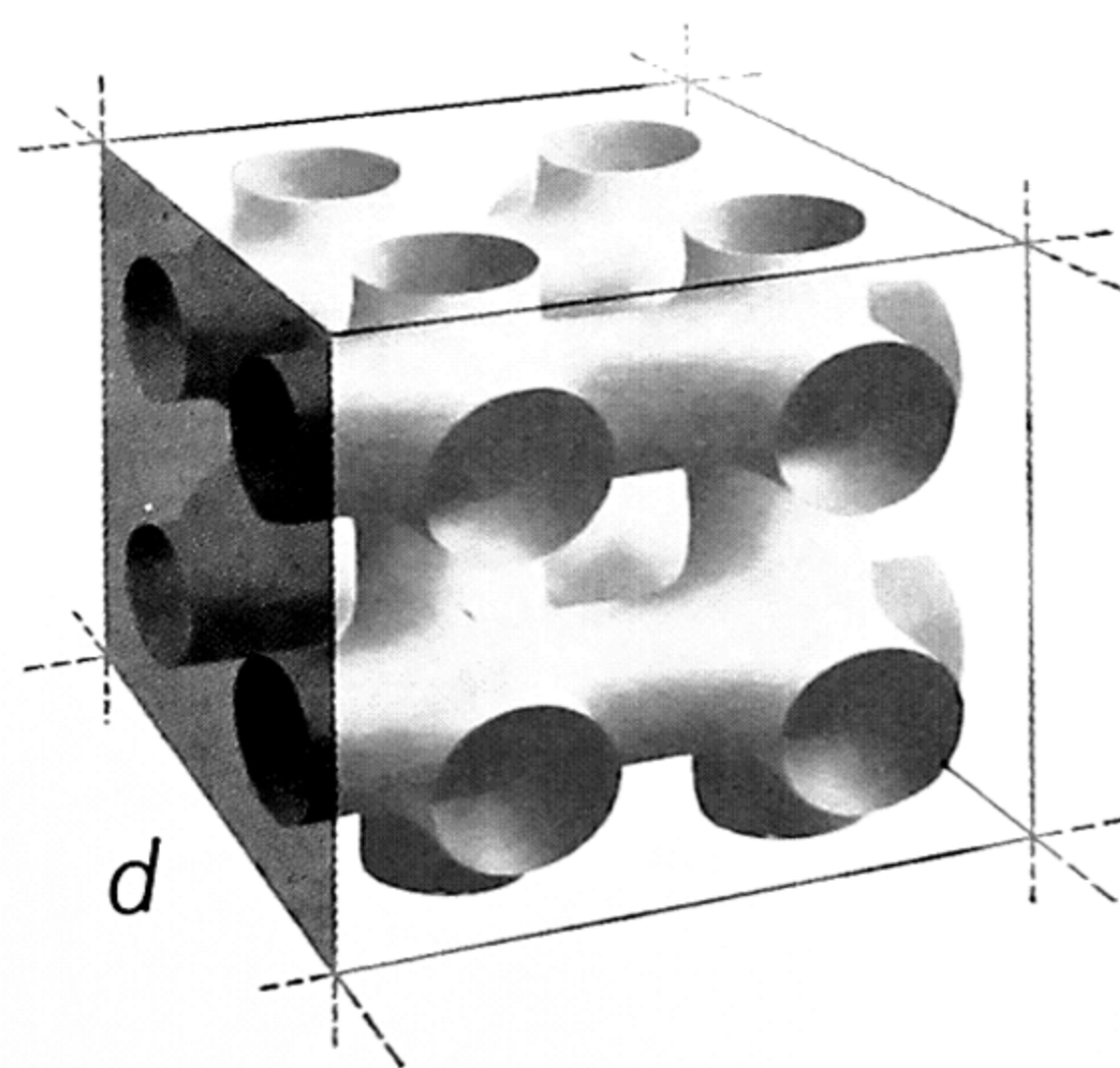
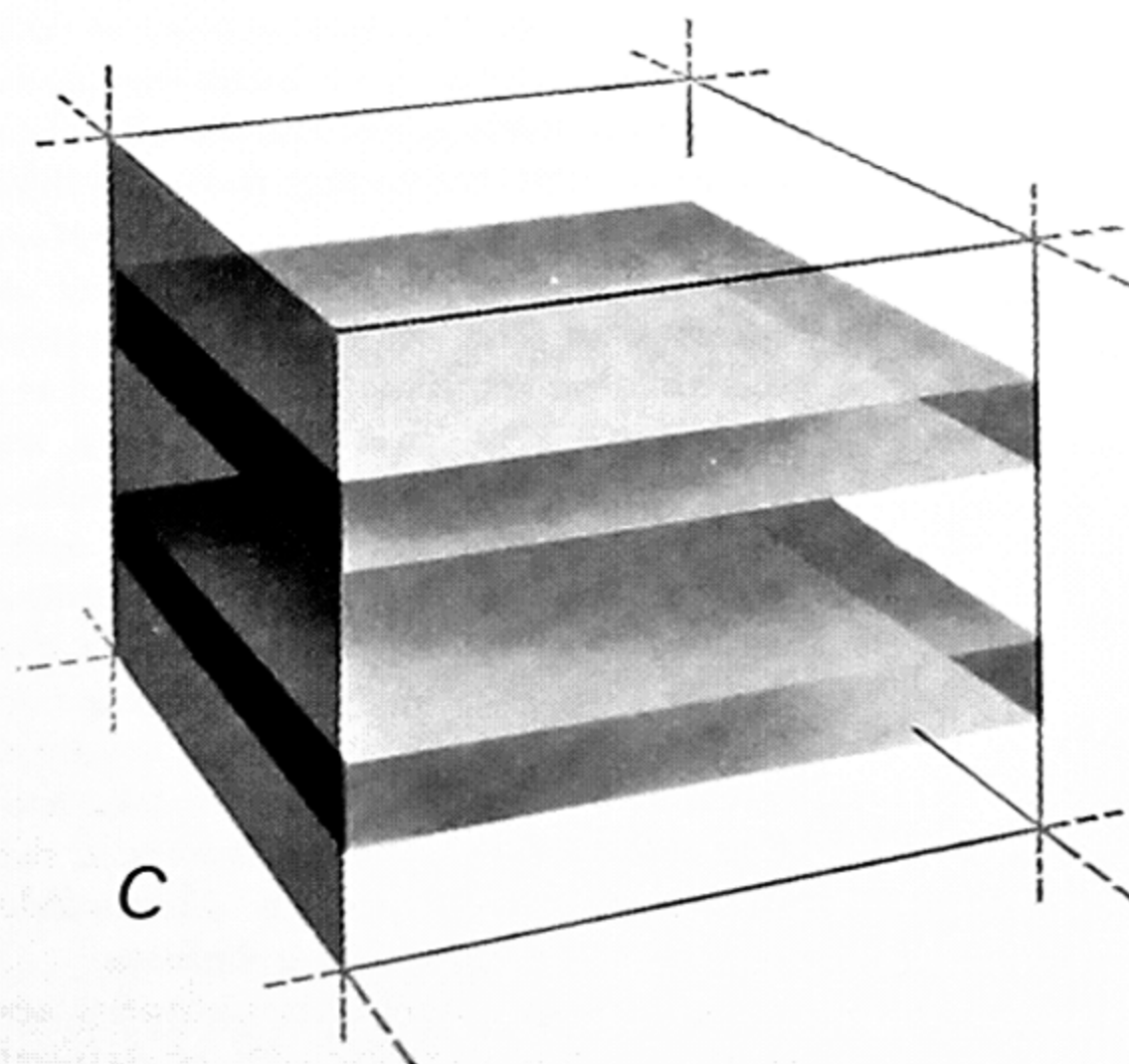
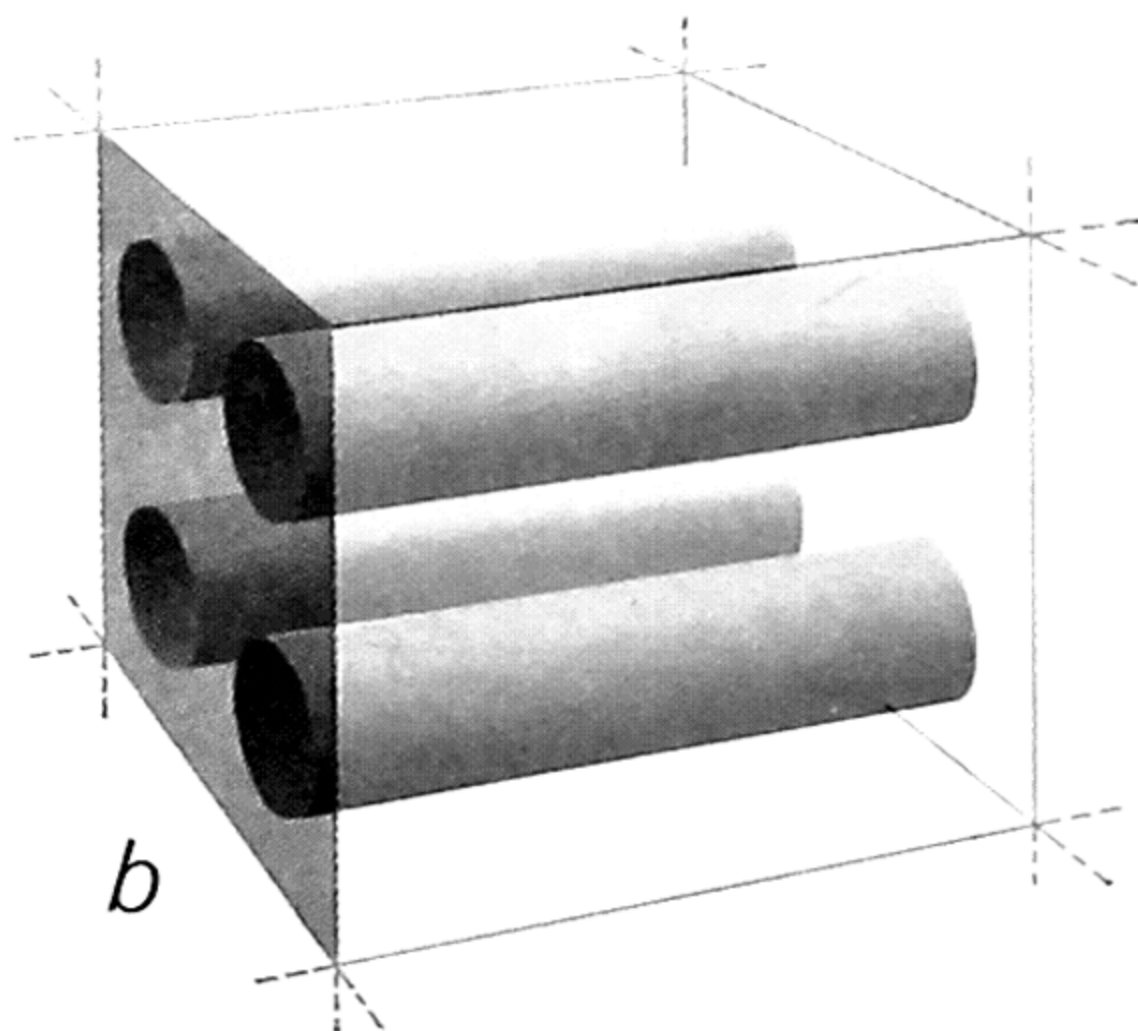
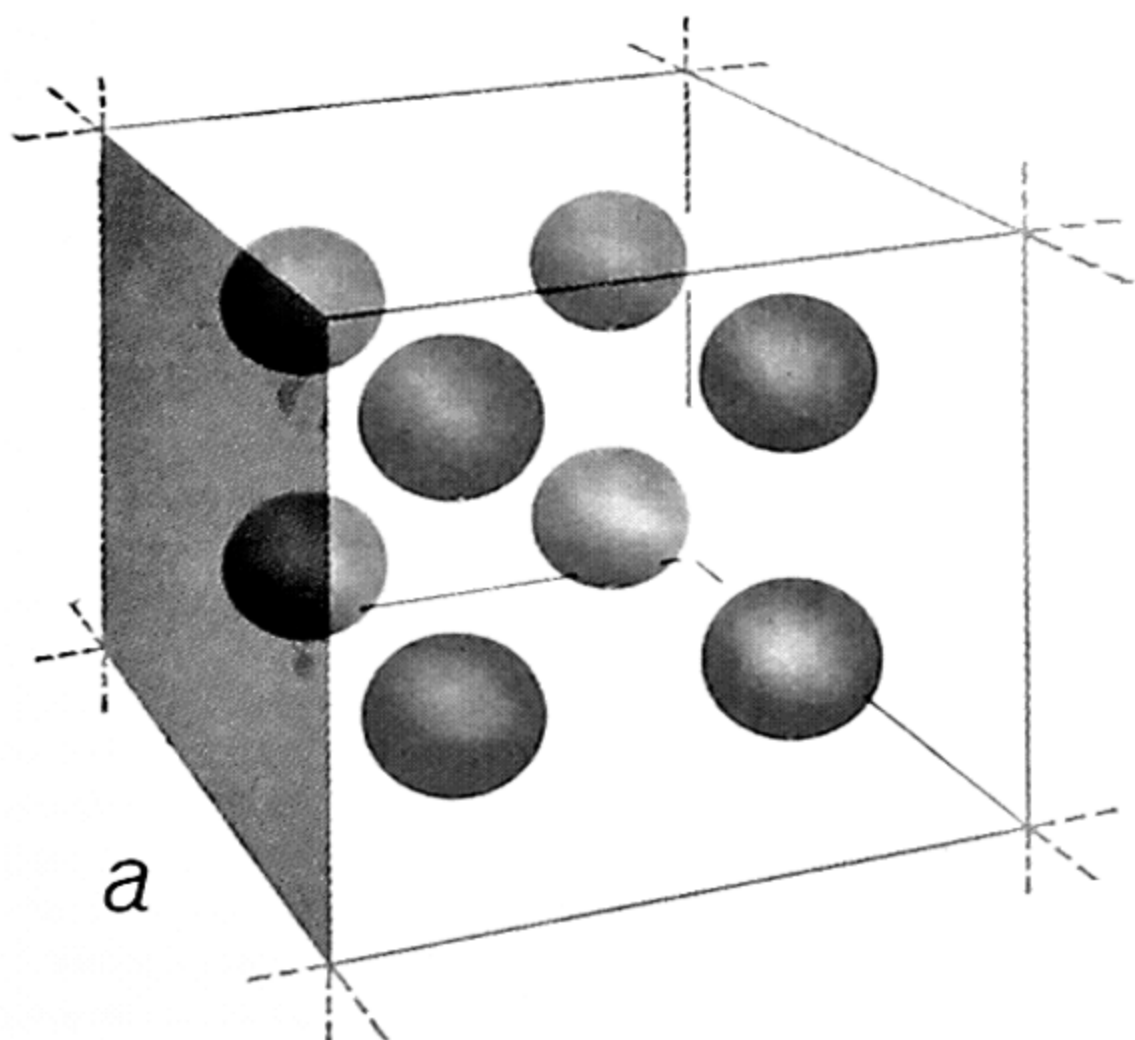
The quinol clathrates are formed only by molecules that can fit into the four-angstrom cavities of the host crystal. In addition to hydrogen sulfide, H_2S , and sulfur dioxide, SO_2 , these include methyl alcohol (but not its closest chemical relatives, water and ethyl alcohol); formic acid, HCO_2H (but not acetic acid, $\text{CH}_3\text{CO}_2\text{H}$); and acetonitrile, CH_3CN (but not any other nitrile). Decisive proof that the clathrate formation depends on molecular form rather

than on chemical bonding is indicated by the fact that even the inert gases argon, krypton and xenon are capable of forming stable quinol complexes.

Quinol is by no means unique in its ability to form inclusion compounds. Hundreds of substances that combine in this way with solvents are known, and it appears that a large fraction of them are clathrates. Various inorganic salts can also form clathrates. An example is the complex of nickel cyanide ammine with benzene.

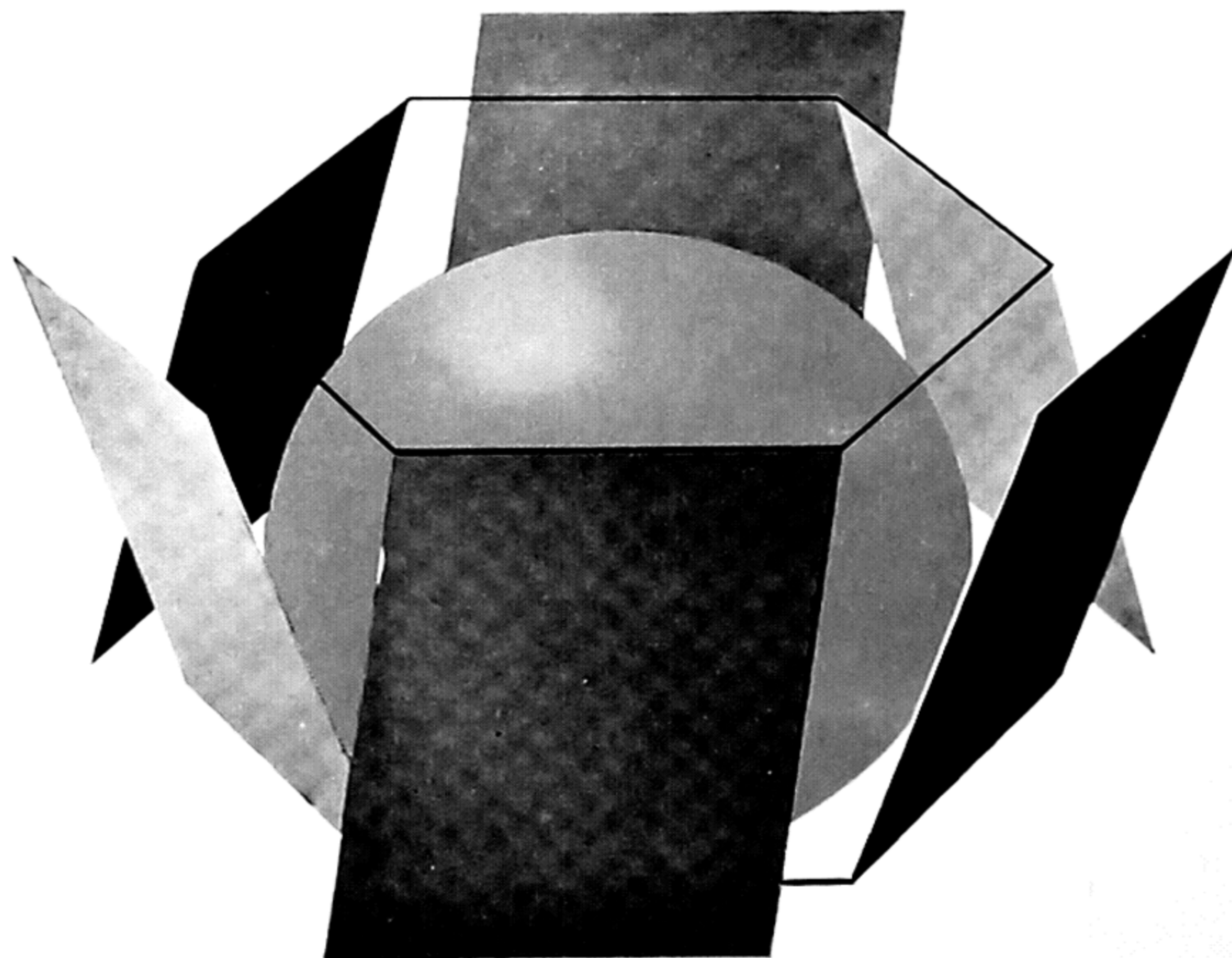
Many substances soluble in water contain "water of crystallization" when they crystallize out of solution. Called hydrates, most of these substances are not clathrates but contain water molecules attached by the usual chemical bonds. The substances with which water does form true clathrates are largely gases or low-boiling-point liquids, and the complexes are generally known as gas hydrates.

Such hydrates fall into two groups. The first usually contains six guest molecules combined with 46 water molecules; the second, one guest molecule for each 17 water molecules. The guests in the first group are small molecules such as chlorine, bromine, sulfur dioxide, hydrogen sulfide, methane, ethane, methyl chloride or methylene chloride. (The chemical formulas for these eight molecules are respectively Cl_2 , Br_2 , SO_2 , H_2S , CH_4 , C_2H_6 , CH_3Cl and CH_2Cl_2 .) The guests in the second group of gas hydrates are slightly larger molecules, such as chloroform, ethyl chloride, methyl iodide, difluorobromochloromethane or propane. (The formulas are CHCl_3 , $\text{C}_2\text{H}_5\text{Cl}$, CH_3I , CF_2BrCl and C_3H_8 .) The gas hydrates of both groups are simple crystals with low melting points. They are generally prepared by forcing the gas

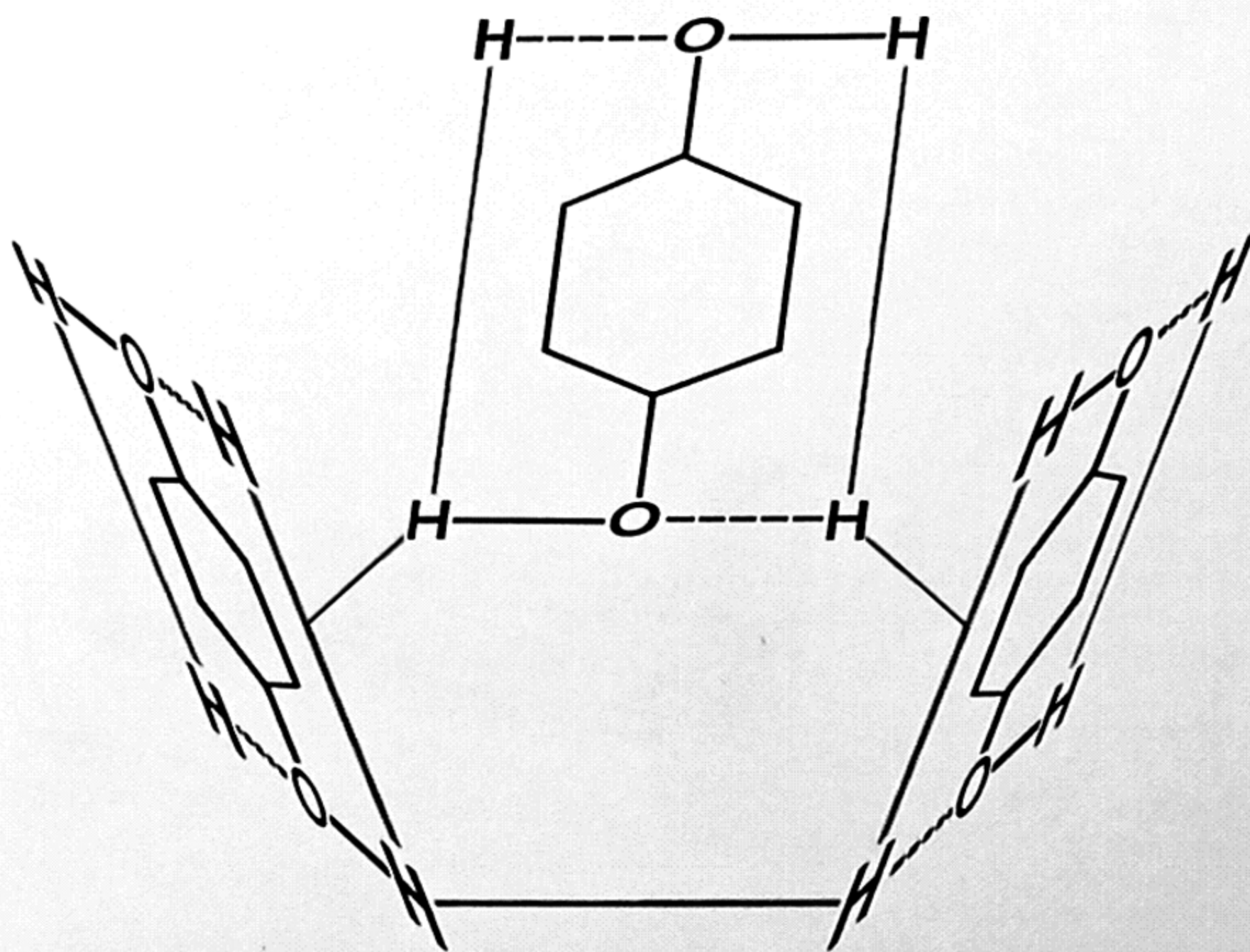


INCLUSION COMPOUNDS are known in six different forms, depending on the architecture of the "host" structures and the shapes of the cavities they enclose. Clathrates (*a*) are produced when "guest" molecules fit into separate spherical chambers within a crystal lattice. Canal complexes (*b*) are formed when the host is a crystal lattice having tubular cavities. Layer complexes (*c*) are

crystals with alternating layers of guest and host. Molecular sieves (*d*) are formed by crystals that contain interconnecting chambers and passageways. Intramolecular hollow space complexes (*e*) are formed when the host is a large molecule containing a concavity. Linear polymer complexes (*f*) are produced when guest molecules fit into the tube formed by a pipelike host molecule.



QUINOL COMPLEX is a clathrate. The quinol molecules (*rectangles*) fit together in sets of six to form an interlocking network. The guest molecules enclosed by six quinol molecules are seen in color. The cavity holding these guests is roughly spherical and about four angstrom units in diameter. No chemical bonds connect the guest molecules with their host.



QUINOL MOLECULES in a quinol complex are linked by hydrogen bonds between oxygen atoms (*O*). The cavities of the complex are bounded by hydroxyl groups (*OH*) and by the benzene rings of the quinol molecules. The diagram shows three molecules and their bonds.

to dissolve in water under pressure and then chilling the solution until it freezes.

X-ray crystallographic studies by M. von Stackelberg and his collaborators at the University of Bonn, by Walter F. Claussen at the University of Illinois and by Linus Pauling and Richard E. Marsh at the California Institute of Technology have shown that in these clathrates the water molecules are linked together, through hydrogen bonds, mostly in rings of five molecules rather than the rings of six molecules typical of ordinary ice. These five-molecule rings are joined together to form dodecahedrons (12-sided figures). Since space cannot be filled completely by any packing arrangement of dodecahedrons, some interstitial spaces must remain. It is these that hold the guest molecules [see illustration on opposite page].

One major application of gas-hydrate formation is now under study by the Koppers Company, Inc. It has long been known that fresh water can be obtained from sea water by partial freezing. The ice crystals that form are pure water; the salt remains behind in the brine. A limitation of this process as a practical means of desalination is the cost of refrigeration. This can be reduced by creating a gas-hydrate "ice" that forms at a higher temperature than does ordinary ice. In the Koppers process propane is added to cold sea water, creating the hydrate $C_3H_8 \cdot 17H_2O$, which freezes at 42 degrees Fahrenheit. The hydrate crystals are filtered off and warmed under pressure. When the crystals melt, the propane separates as an immiscible liquid, leaving pure water behind.

Very similar to the gas hydrates are the hydrates of certain salts containing hydrocarbon chains linked to sulfur or nitrogen atoms. They are known technically as trialkylsulfonium and tetraalkylammonium salts. One example is the compound $2(n-C_4H_9)_3S^+F^- \cdot 40H_2O$. The crystal lattice formed by this substance is similar to that in the gas hydrate containing 46 water molecules, except that six of the water molecules are replaced by the two ions of sulfur and two of fluorine, and the cavities are filled by the six hydrocarbon chains (C_4H_9) rather than by six gas molecules.

Another example is the hydrate with the formula $2(i-C_5H_{11})_4N^+F^- \cdot 76H_2O$. This is a surprising substance, 68 per cent water, that remains an ice until heated to 88 degrees F. Proteins also form highly hydrated crystals, containing as much as 90 per cent water, or thousands of water molecules for each giant molecule of protein. It may well be

that the protein holds these water molecules in modified clathrate structures not unlike those in the hydrates of the sulfonium and ammonium salts.

Molecules in Tubes

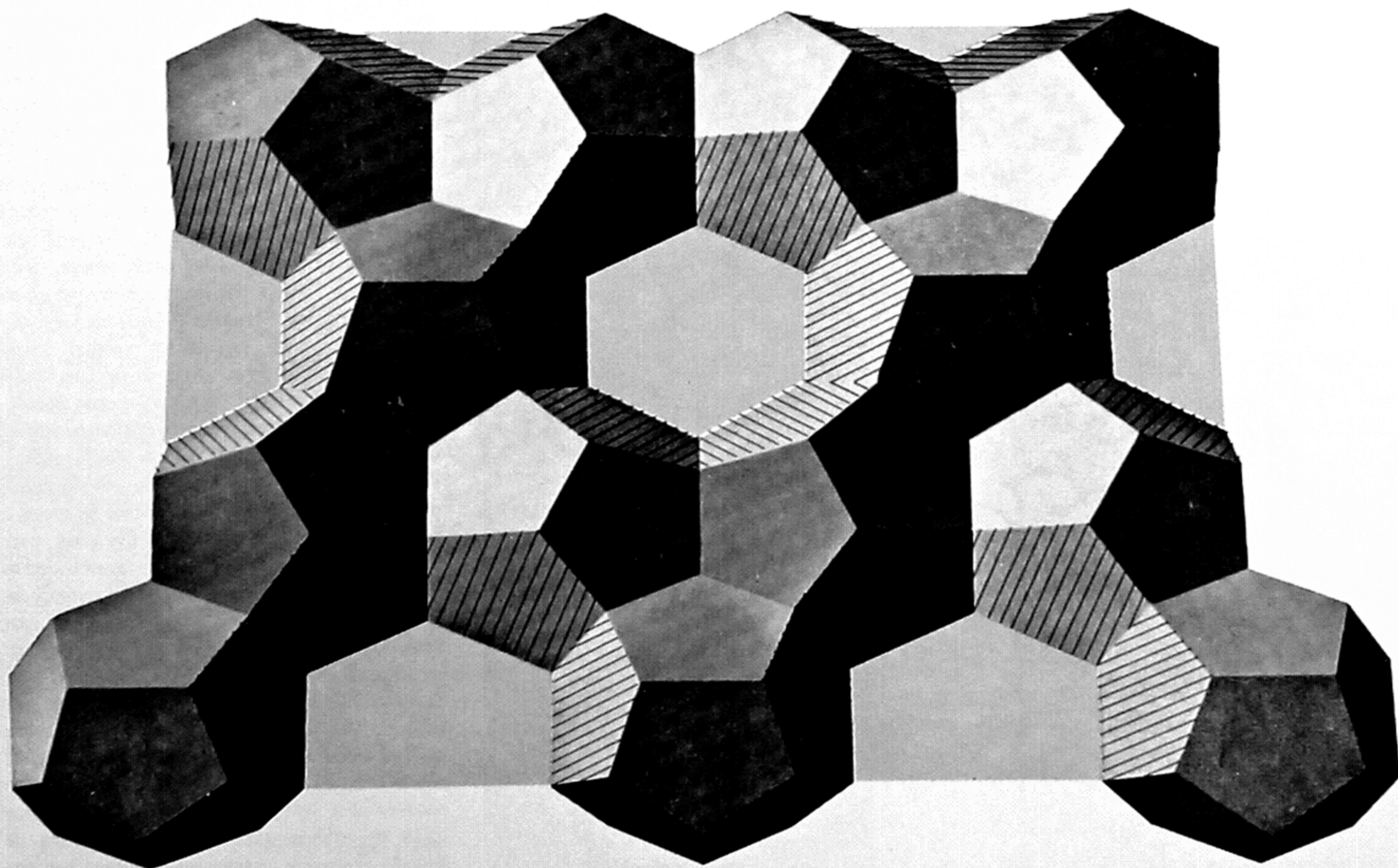
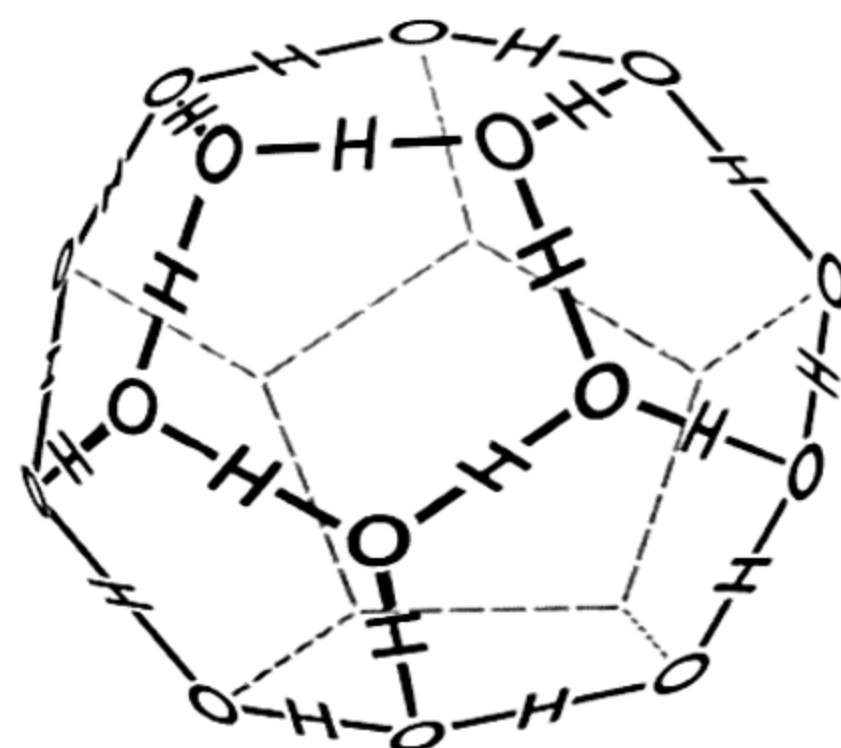
A rather different sort of structure has been found in the crystalline inclusion compounds formed by urea, thiourea, deoxycholic acid, starch and a few other substances. Here the cavities in the host crystal consist of long tubes in which the guest molecules lie end to end. This type of inclusion compound is called a canal complex. The best known of the canal complexes are those formed by urea; they were discovered in 1940 by the German chemist M. F. Bengen. The extraordinary specificity of urea in complexing almost exclusively with straight-chain hydrocarbons and their derivatives attracted the attention of several research groups in the 1940's, and the canal-complex structure was established at the end of the decade by

the German chemist W. Schlenk, Jr., and by Albert E. Smith of the Shell Development Company in the U.S.

In the urea canal complexes, the molecules of urea, $\text{CO}(\text{NH}_2)_2$, are held together by hydrogen bonds between the nitrogen and oxygen atoms. The interconnected urea molecules are arranged in much the same way as the wax in a honeycomb, leaving long tubular cavities in which the guest molecules reside [see illustration on next page]. The inside diameter of the canals is about five angstrom units, which is just the right size to accommodate the straight-chain hydrocarbons, the fatty acids or derivatives of both. Any such substance having a chain length of more than six carbon atoms will complex with urea at room temperature, and some substances that are only three or four atoms long will complex at low temperatures.

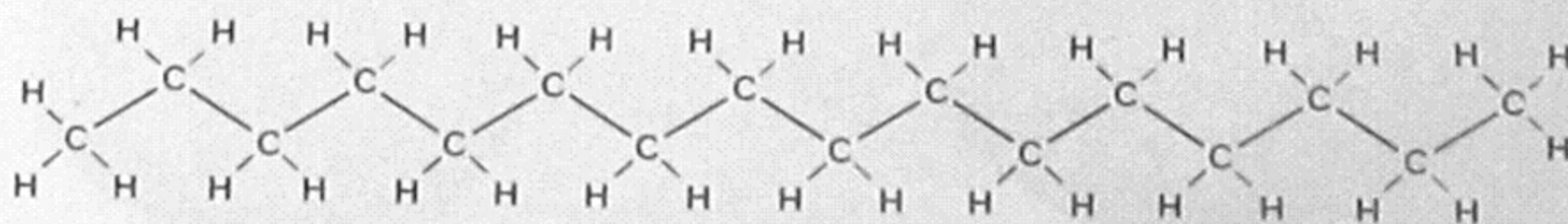
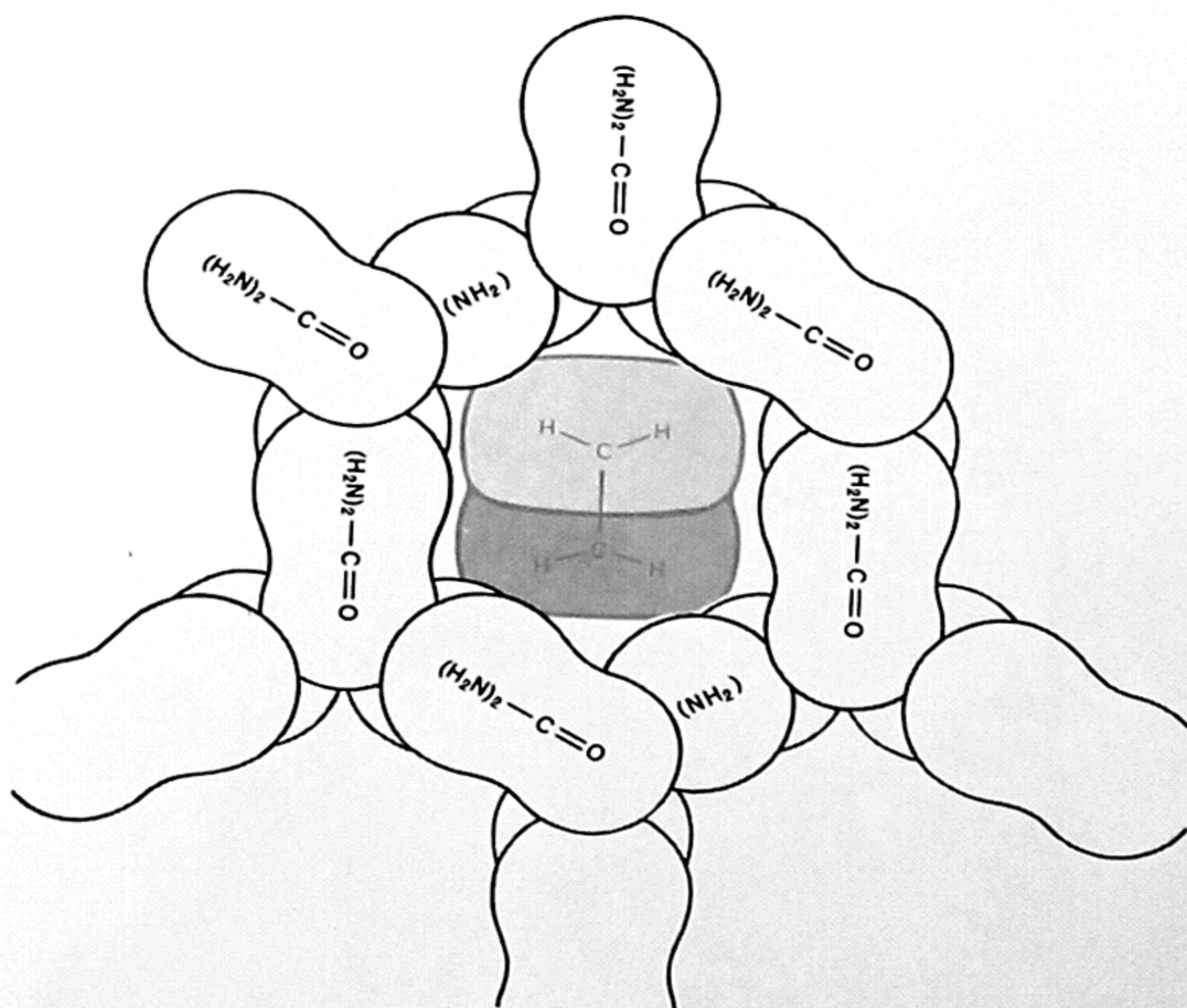
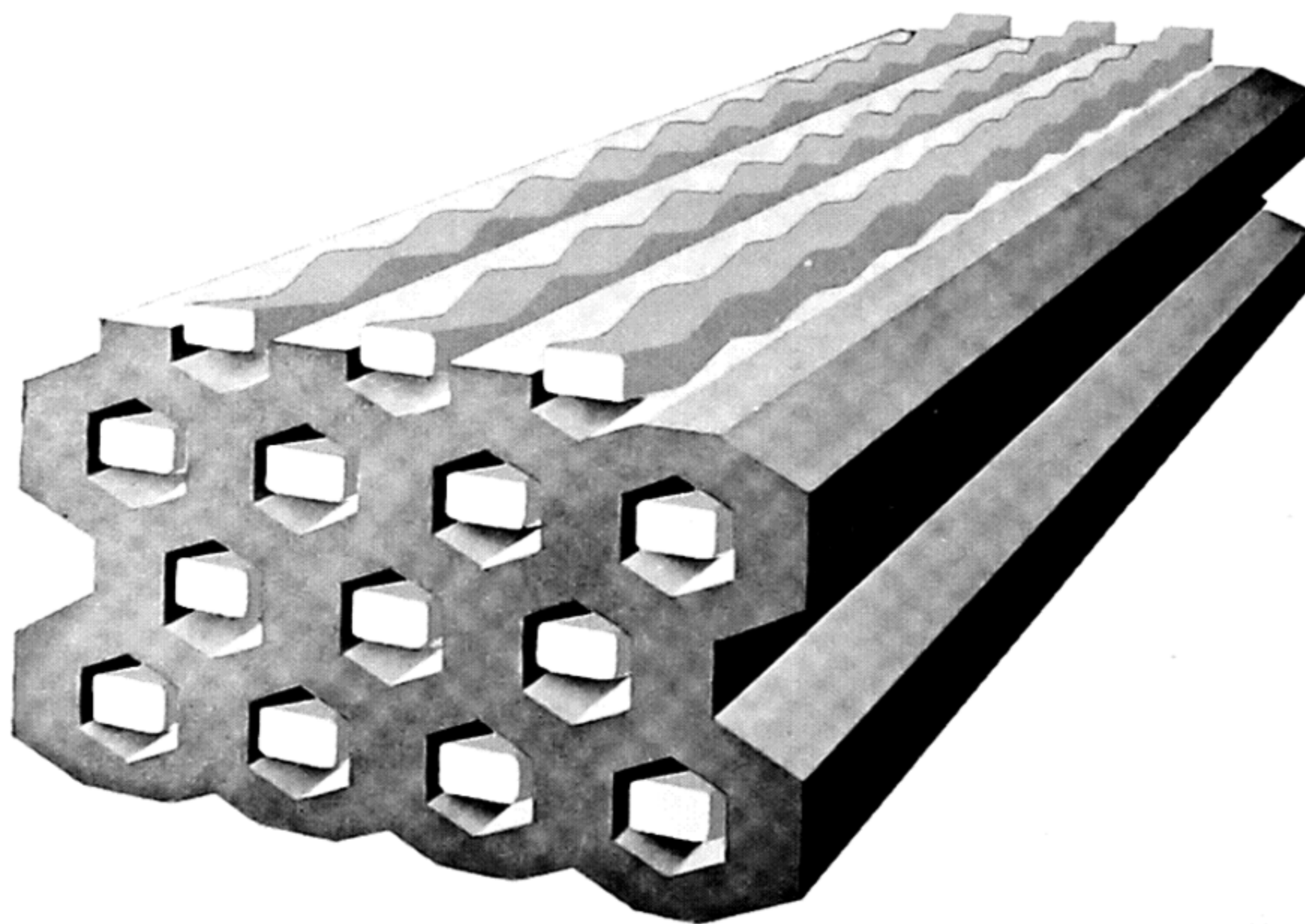
Thiourea, $\text{CS}(\text{NH}_2)_2$, the sulfur analogue of urea, also forms canal complexes. In these honeycombs, however,

the canal diameter is seven angstroms instead of five because of the greater size of the sulfur atom. The substances that can pack into these seven-angstrom holes, thereby forming thiourea canal complexes, constitute a rather capricious and unpredictable group. It includes some of the simpler ring-shaped hydrocarbons (cycloparaffins) as well as a number of polymethylated and poly-



GAS HYDRATES are clathrates composed of water molecules and molecules of a gas or a low-boiling-point liquid. The water molecules link together through hydrogen bonds in rings of five. These

rings are combined in dodecahedrons like the one at top right. Since dodecahedrons cannot be made to pack together precisely, cavities (color) are formed into which the guest molecules fit.



CANAL COMPLEX is formed when guest molecules lie end to end in tubes within a crystal. A canal complex with urea molecules as the host is shown in the drawing at top. The guests are in color. The molecules of urea, seen in the middle drawing, are held together by hydrogen bonds between nitrogen and oxygen atoms to form a honeycomb-like structure. Linear hydrocarbon chains, in bottom drawing, can fit into the cavities of the honeycomb.

chlorinated hydrocarbons—that is, hydrocarbons with methyl groups or chlorine atoms projecting from the central chain. But it does not include the straight-chain, paraffin-like hydrocarbons that complex so well with urea.

Several dozen patents describing practical uses of the urea and thiourea complexes have already been issued, and the list is growing steadily. The majority are based on the ability of inclusion compounds to sort out molecules on the basis of shape and thus effect separations of mixtures. For example, the paraffinic constituents of lubricating oil, which have an undesirable tendency to solidify in cold weather, can be removed by mixing the oil with urea and allowing the urea-paraffin complex to form. Conversely, many cycloparaffins (naphthenes and terpenes) can be separated from gasoline stocks, or from the essential oils of plants, by treatment with thiourea. The canal complexes have also attracted attention as a means of storing easily oxidized materials, such as unsaturated fatty acids or vitamin A. While they are held as guest molecules in canal complexes these sensitive substances cannot be attacked by atmospheric oxygen. Other patents claim the use of urea complexing to convert sticky, semisolid detergents into easily handled powders, and similar uses of thiourea complexing to facilitate handling of chlorinated insecticides such as chlordane.

Although the guest molecules in a canal complex are inaccessible to attack by external reagents, they are still capable of reacting with each other. Several years ago it occurred to some of us at the General Electric Research Laboratory that such reactions might yield rather unusual products, since the walls of the canal would impose severe limitations on the ways in which the reacting molecules could approach each other. Conceivably, in a reaction in which small monomer molecules are linked to form a polymer, there might be only one way in which the reacting monomers could join together, and this would result in a polymer with a very orderly, or stereospecific, structure. Investigations in our laboratory by the author and Dwain M. White showed that this was indeed the case. Using brief irradiation with a one-million-electron-volt beam of electrons as a means of initiating polymerization within the canals, we were able to polymerize such monomers as vinyl chloride and butadiene within their urea hosts and 2,3-dimethylbutadiene, 2,3-dichlorobutadiene and 1,3-cyclohexadiene in their thiourea complexes. In all these cases the polymers

obtained were highly crystalline and had high melting points, in sharp contrast with normal polymers obtainable by irradiation of the liquid monomers, thus showing the stereospecific polymerization had occurred. The four-carbon monomers joined together exclusively at the terminal carbon atoms to produce what are called 1,4-*trans* addition polymers [see illustration below]. Evidently canal complexes can act as molecular templates, the smooth, straight holes in a canal host giving rise to smooth, straight molecules of polymer.

Inclusion-compound formation is by no means restricted to crystals. Even in cases where the cavities in the host are built up from separate molecules, there are often indications of host-guest relations in concentrated solutions, comparable to those occurring in the crystal. For example, the formation of micelles (aggregates of soap molecules that produce opalescence in concentrated soap solutions) can be inhibited by the addition of urea, implying association of the urea molecules with the hydrocarbon groups of the soap. It has also been suggested that the inactivation, or denaturation, of proteins by strong urea solutions, results from inclusion-compound formation with the hydrocarbon groups of the protein. Localized canal-complex formation probably also accounts for the action of bile in increasing the solubility of fats in the digestive tract. One constituent of bile, deoxycholic acid, is noted for the stability of the canal complexes that it forms with simple organic molecules.

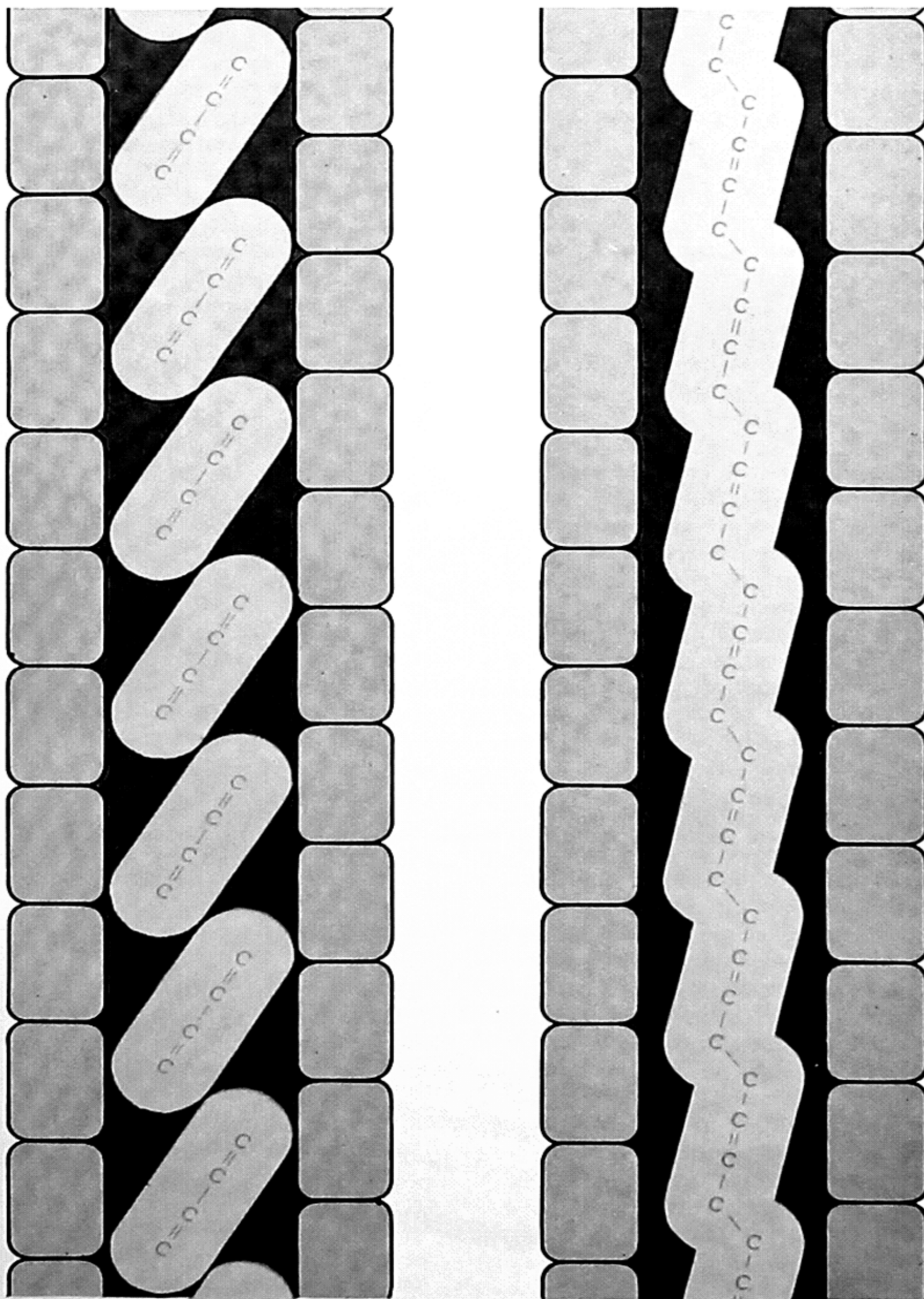
There is abundant evidence that hydrate-like "iceberg" layers exist around many ions, inert gases and protein molecules in aqueous solution. Recently Pauling has suggested that the anesthetic action of certain gases of low chemical reactivity (for example, chloroform and cyclopropane) may result from hydrate formation. There is a remarkably close parallel between the anesthetic activity of these substances and the stability of their crystalline hydrates. Xenon, which enters into no combinations except clathrates, is an effective anesthetic. It is proposed that "microcrystals" of gas hydrate form on certain proteins in the brain and inhibit their functioning.

The Starch-Iodine Reaction

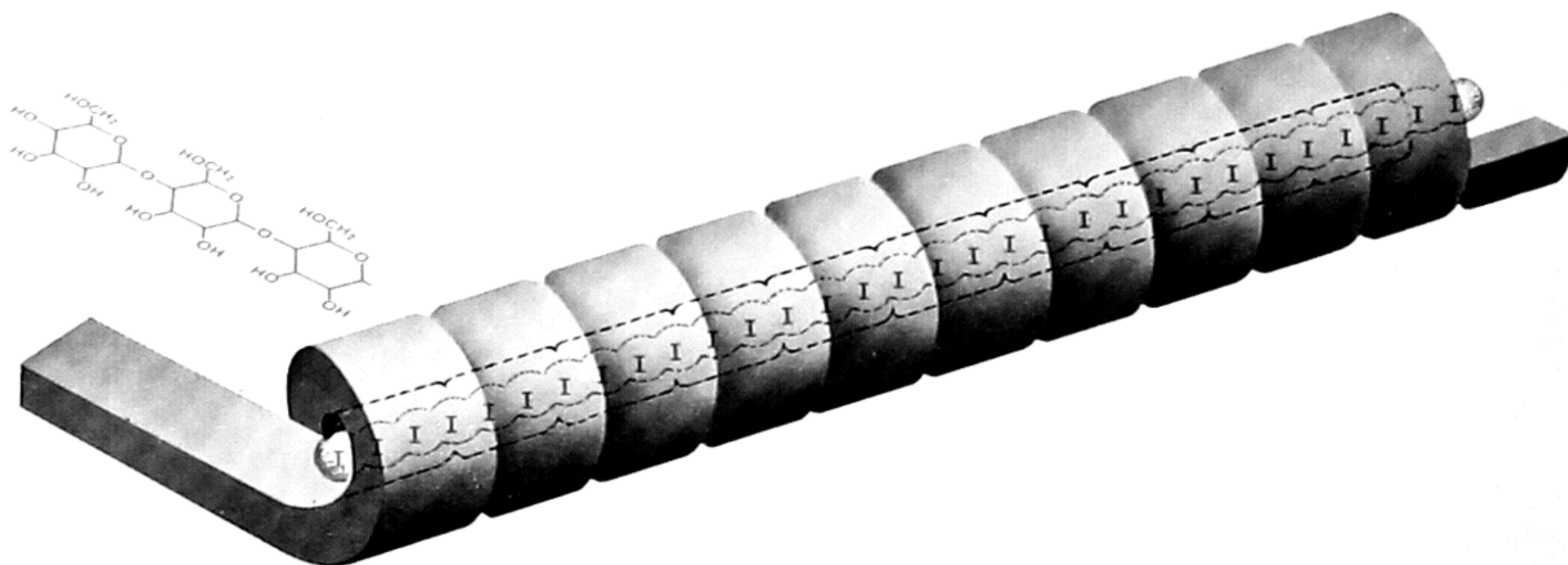
A remarkable family of inclusion compounds can be formed with iodine as the guest substance. It has been known since early in the 19th century that starch and iodine form an intensely colored, blue-black complex. That the complex is a

soluble inclusion compound of the canal type has been known for only a decade or so, mainly as a result of work by Karl Freudenberg and Friedrich D. Cramer at the University of Heidelberg and by Dexter French and Robert E. Rundle at Iowa State University. Their studies show that the iodine atoms are joined together to form a long, straight polyiodide chain and that the starch molecule, which is a linear polymer composed of glucose units, forms a spiral wrapping

around these chains [see top illustration on page 692]. Neither the helical configuration of the starch molecule nor the chain of iodine atoms is stable except in inclusion structures of this general type. The starch helix will form, however, around guests other than iodine, and the iodine chain will occupy hosts other than starch. For example, the free fatty acids that often occur along with starch in plants can make the starch take the shape of a hollow helix. Moreover, sev-

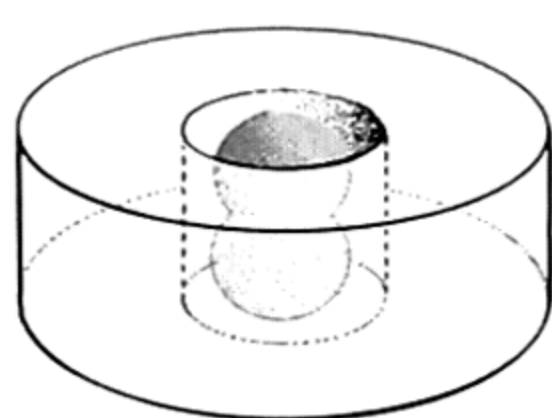


POLYMERIZATION of the guest molecules in a urea canal complex is obtained by irradiating the inclusion compound with X rays or high-energy electrons. Since the shape of the cavity acts as a template, the four-carbon monomers (butadiene) in drawing at left can join only at their terminal atoms. At right is the orderly polymer (polybutadiene) thus produced.



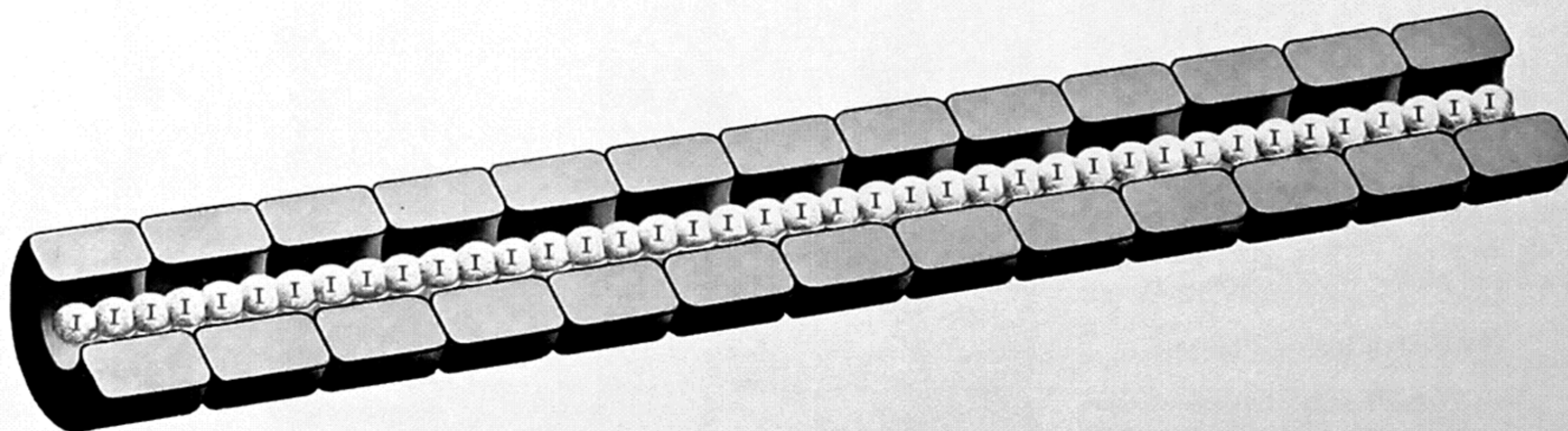
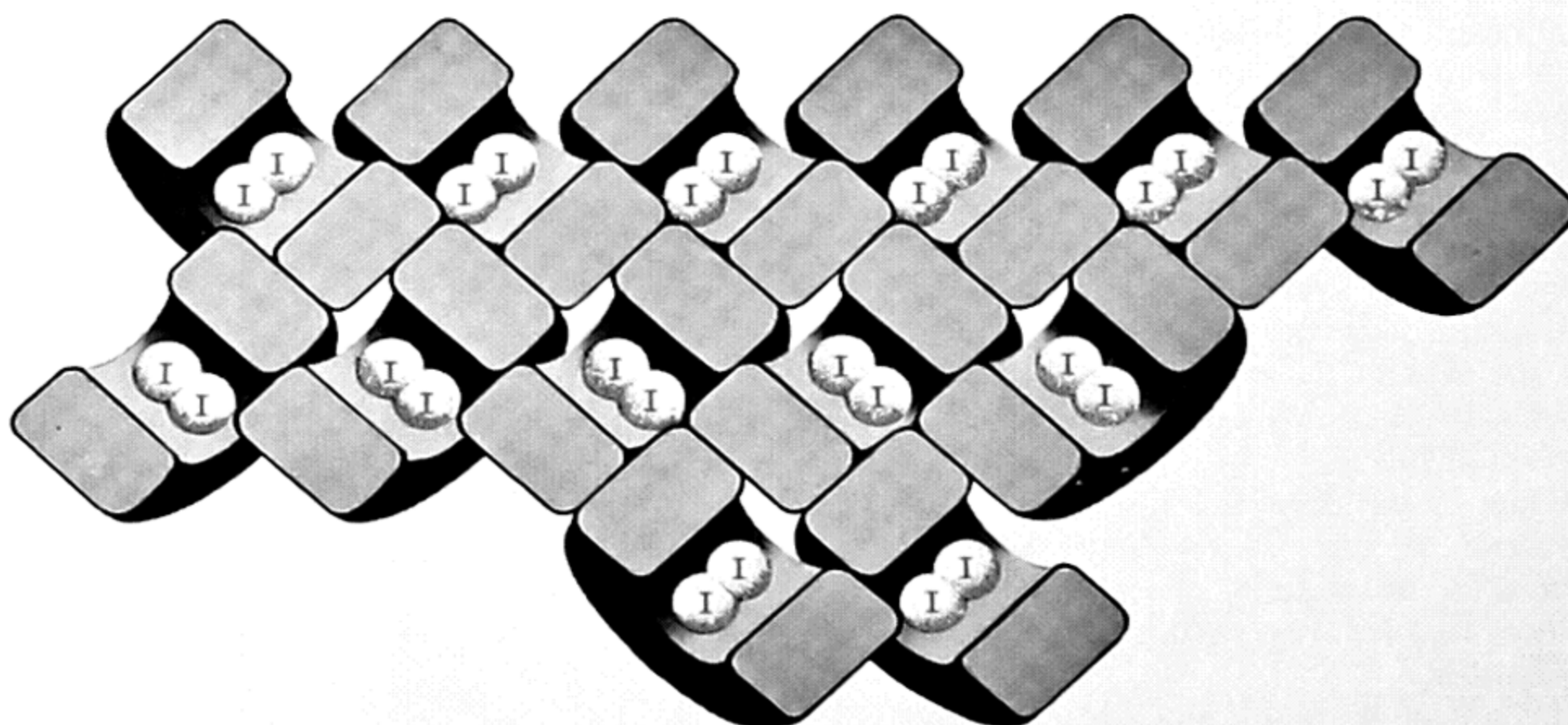
CYLINDRICAL STRUCTURE in a starch-iodine complex results when a polyiodide chain is enclosed in the cavity formed by a

helical molecule of starch. Each turn of the helix contains six glucose units. The guest molecule is a linear chain of iodine atoms.



CRYSTAL PACKING ARRANGEMENT in a molecular iodine-cyclohexadextrin clathrate results when the doughnut-shaped

molecules of cyclohexadextrin enclose separate iodine molecules. This complex can also exist as separate molecules in solution.



TUBULAR PACKING ARRANGEMENT in an iodine-cyclohexadextrin complex results when polyiodide chains are enclosed in

the cavities of cyclohexadextrin molecules lying side by side. The complex can also exist as separate molecules in solution.

eral other canal hosts, such as cyclodextrin, polyvinyl alcohol and barbituric acid, can stabilize the linear polymers of iodine.

It happens that the behavior of electrons in a linear polyiodide chain resembles that of the electrons in a row of metal atoms. Some of the electrons are free to move back and forth along the chain; this accounts for the intense absorption of light by the starch-iodine complex. The electrons are not free to move perpendicular to the chain, however. This property has been utilized in the preparation of Polaroid filters for polarizing light. In these filters a sheet of the polymer polyvinyl alcohol is treated with iodine to produce the linear polyiodide inclusion compound and is then stretched to orient the iodine chains. Such sheets will absorb the component of light polarized parallel to the direction of stretch and allow only the component that is polarized at right angles to come through.

Perhaps the most striking of the soluble substances that form inclusion compounds are the cyclodextrins, which have been extensively studied by Cramer. The cyclodextrins are rigid, doughnut-shaped molecules formed by joining glucose units together in rings. Thus they are the cyclic analogues of amylose (starch), which contains the same glucose units joined together in long chains. Three cyclodextrins are known. These contain six, seven and eight glucose units and have internal diameters of about six, eight and 10 angstroms respectively. All three cyclodextrins combine avidly with any molecule that can fit wholly or partially into these holes. Most of the complexes can also be crystallized out of solution. In the solid state they resemble either the clathrates or the canal complexes. Both forms can be obtained with the six-membered cyclodextrin acting as host to iodine. The clathrate contains separate iodine molecules (I_2) and has the characteristic red-brown color of ordinary iodine solutions. The canal complex, containing polyiodide chains, is blue-black.

Tailor-made Hosts

The applications of inclusion compounds for separating mixtures of molecules of different shape, or as templates for guiding the direction of chemical reactions, all depend on the selectivity of the host-guest interaction. Inclusion-compound formation does not occur unless there is a fairly good match between the guest molecule and the cavity in the host. This selectivity imposes limitations

on the widespread use of inclusion compounds in these applications. Among the simple synthetic systems only a few dozen inclusion hosts are known, and these cover only a narrow range of cavity shapes.

One approach to the tailoring of host cavities to any desired form was discovered some years ago by Frank H. Dickey at the California Institute of Technology. He found that when silica gel was formed in the presence of certain dyes and subsequently extracted to remove the dye, the remaining gel had a strong and specific adsorptive capacity for the dye that had been present during its formation. Gel made in the presence of butyl orange, $(C_4H_9)_2NC_6H_4N=NC_6H_4SO_3Na$, was a much better adsorbent for that dye than for propyl orange, $(C_3H_7)_2NC_6H_4N=NC_6H_4SO_3Na$, and vice versa. Because of this selectivity in their adsorption power, these gels were called synthetic antibodies.

Silica gel has an irregular three-dimensional network structure in which silicon atoms are connected by oxygen atoms and occupy much the same positions with respect to one another as do the hydrogen-bonded oxygen atoms in water. When the silica-gel network is still incompletely connected, it is flexible enough to conform to the subtle differences in shape presented by different kinds of guest molecule. After the cross-linking of the network is complete and the gel has been dried, the guest can be extracted, leaving behind a phantom of its shape. The synthesis of these tailor-made hosts is in some ways just the inverse of a canal-complex polymerization. There the cavity in the complex was used as a template for guiding the polymerization of the guest molecules. Here the guest molecules are used as templates for guiding the polymerization of the host.

Inclusion Systems in Nature

Half a century ago the German chemist Emil Fischer proposed that shape-dependent, "lock and key" interactions might explain the extraordinary selectivity of the chemistry that takes place in living cells. The basis for this idea can be described as follows. Chemists had found that when they determined the structure of relatively simple organic molecules, they could explain their reactions in terms of reactive groupings of atoms ("functional groups") in the molecules. In nonbiological systems, that is, in the chemist's flasks, compounds carrying the same functional groups reacted very much the same way.

Thus a reagent that combined with the hydroxyl group in one alcohol would generally react with all alcohols; and the adsorbent that could remove one amine from solution was generally capable of adsorbing any amine. In living systems, however, this was not the case. Each individual alcohol, amine and acid had its own set of transformations. Evidently something more than just the nature of the functional groups was involved. The only reasonable explanation was that this extra something was the over-all shape of the molecule. Such shape-specificity is now readily demonstrable for a wide variety of biochemical processes—for example, in antigen-antibody interactions and in the hundreds of reactions that are catalyzed by enzymes. If it were not for the sensitivity to molecular shape of these processes, living organisms would never be able to accomplish the task of assembling their amazingly intricate molecular structures from complex mixtures of small molecular fragments. In order to account for this dependence on molecular shape, it has long seemed a reasonable hypothesis that much of the chemistry of the cell proceeds with the help of hand-in-glove-fitted structures.

In almost all cases the host molecules involved in these shape-dependent biological processes are proteins. Thus the proteins probably represent by far the largest and most important group of substances that form inclusion compounds. Unfortunately there is still no case in which the molecular structure of a substrate-plus-protein complex has been determined, so that the details of the combination are still unknown. There is a prevalent feeling, however, that weak chemical bonds, such as hydrogen bonds and electrostatic attractions, are probably quite often involved, in addition to the binding due to nonlocalized dispersion forces that arise from the precision of the shape-matching. In other words, the substrate-protein complexes may have closer synthetic analogues in the alkyl-sulfonium and alkylammonium salt hydrates, where the host-guest binding results from electrostatic forces in addition to shape-matching, than in the gas hydrates or urea canal complexes, where the binding results from shape-matching alone. In any event, the discovery and detailed understanding of the synthetic inclusion compounds have provided powerful support for the hypothesis that such lock-and-key structures can indeed account for biochemical specificity, and hence that inclusion structures play a central role in the functioning of living organisms.

The Author

JOHN F. BROWN, JR. is an organic chemist at the General Electric Research Laboratory in Schenectady, N.Y. Brown obtained an Sc.B. at Brown University in 1947, did graduate work in physical and organic chemistry at the Massachusetts Institute of Technology, receiving his Ph.D. in 1950, and went to General Electric the same year. From 1956 to 1961 he was manager of the Reaction Studies Unit at General Electric. Brown's interest in enzyme models and techniques for controlling synthetic reactions led him to study inclusion compounds.

Bibliography

- INCLUSION COMPOUNDS. Friedrich D. Cramer in *Reviews of Pure and Applied Chemistry*, Vol. 5, No. 3, pages 143-164; September, 1955.
- A MOLECULAR THEORY OF GENERAL ANESTHESIA. Linus Pauling in *Science*, Vol. 134, No. 3471, pages 15-21; July 7, 1961.
- X-RAY ANALYSIS OF ORGANIC STRUCTURES. S. C. Nyburg. Academic Press Inc. See pages 284-296.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

TELEPHONE SWITCHING

by H. S. Feder and A. E. Spencer

The operation of the network linking the 75 million telephones in the U.S. requires fast and efficient switching. Present electromechanical switching systems may soon be replaced by even faster electronic ones.

Right now you can pick up your telephone and make connection with any of the more than 75 million other telephones in the U.S. On the average 20,000 people throughout the country will have begun calls by the time you have read this far. By dialing a few numbers each caller will set up, in a matter of seconds, a unique pathway to another telephone. The path may be a mile or 3,000 miles long, and creating it can involve the operation of thousands of individual switches. When the call is finished, the path will be dismantled almost instantaneously and its parts will be ready to form new connections.

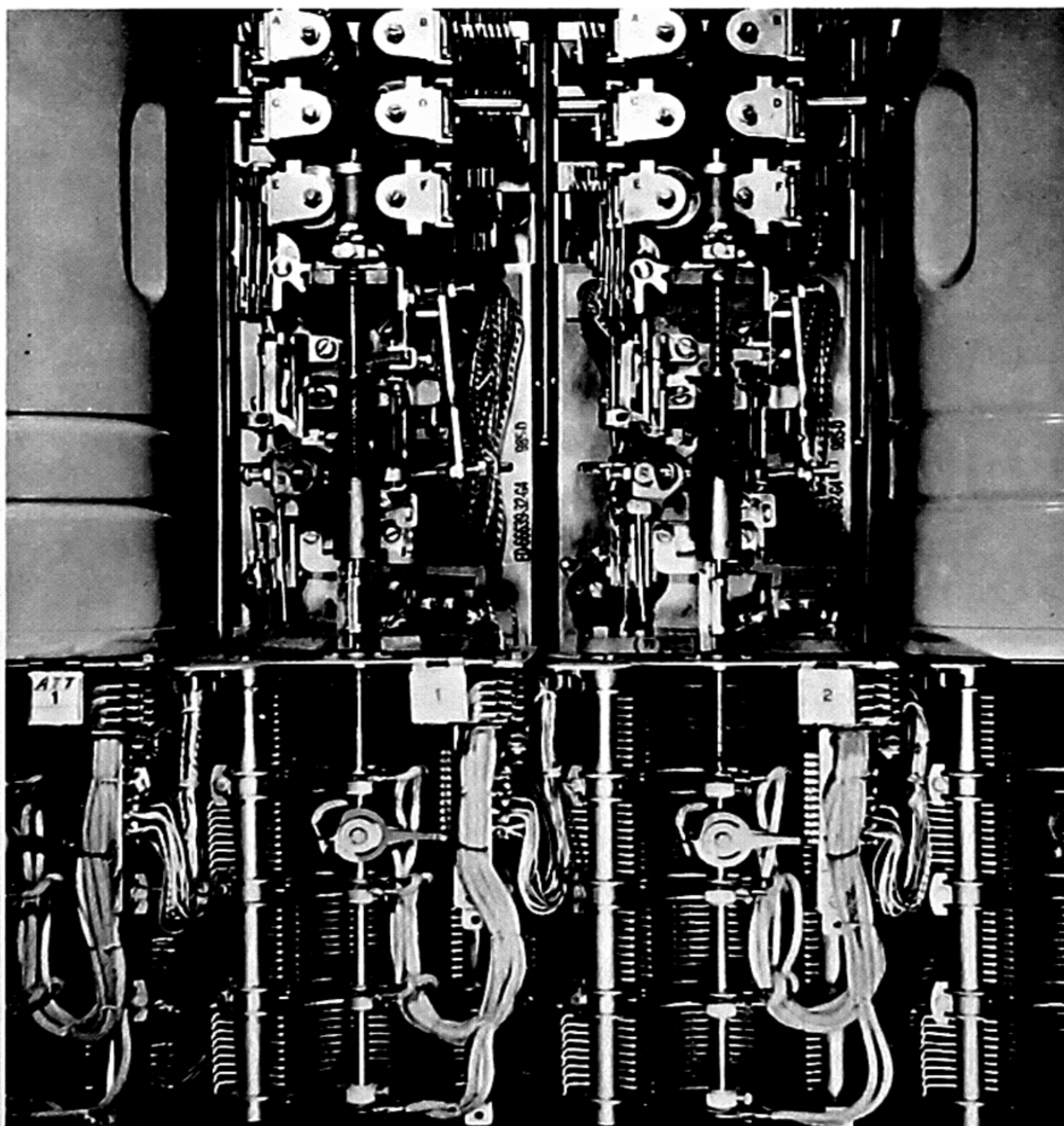
Even to people in the business, when they stop to think about it, the capability of the modern telephone switching system is fantastic. Built on one simple operation—the closing of a switch—these systems literally exceed the bounds of comprehension in their detail. Yet they operate controllably and predictably. Moreover, substantial improvements in speed and flexibility are in prospect.

All this began soon after Alexander Graham Bell patented the telephone in 1876. The telegraph had already posed a problem of electrical switching, but the demands of the simple commercial networks then in existence were modest. The telephone was a different matter: it was an instrument everyone could use. At first few people other than Bell himself had the imagination to think of the telephone as a part of their everyday lives. The new device was sold in pairs, for private communication between a house and barn, say, or a home and a nearby place of business. The early advertisements even warned of a possible lack of privacy if the purchaser connected a third telephone to his system!

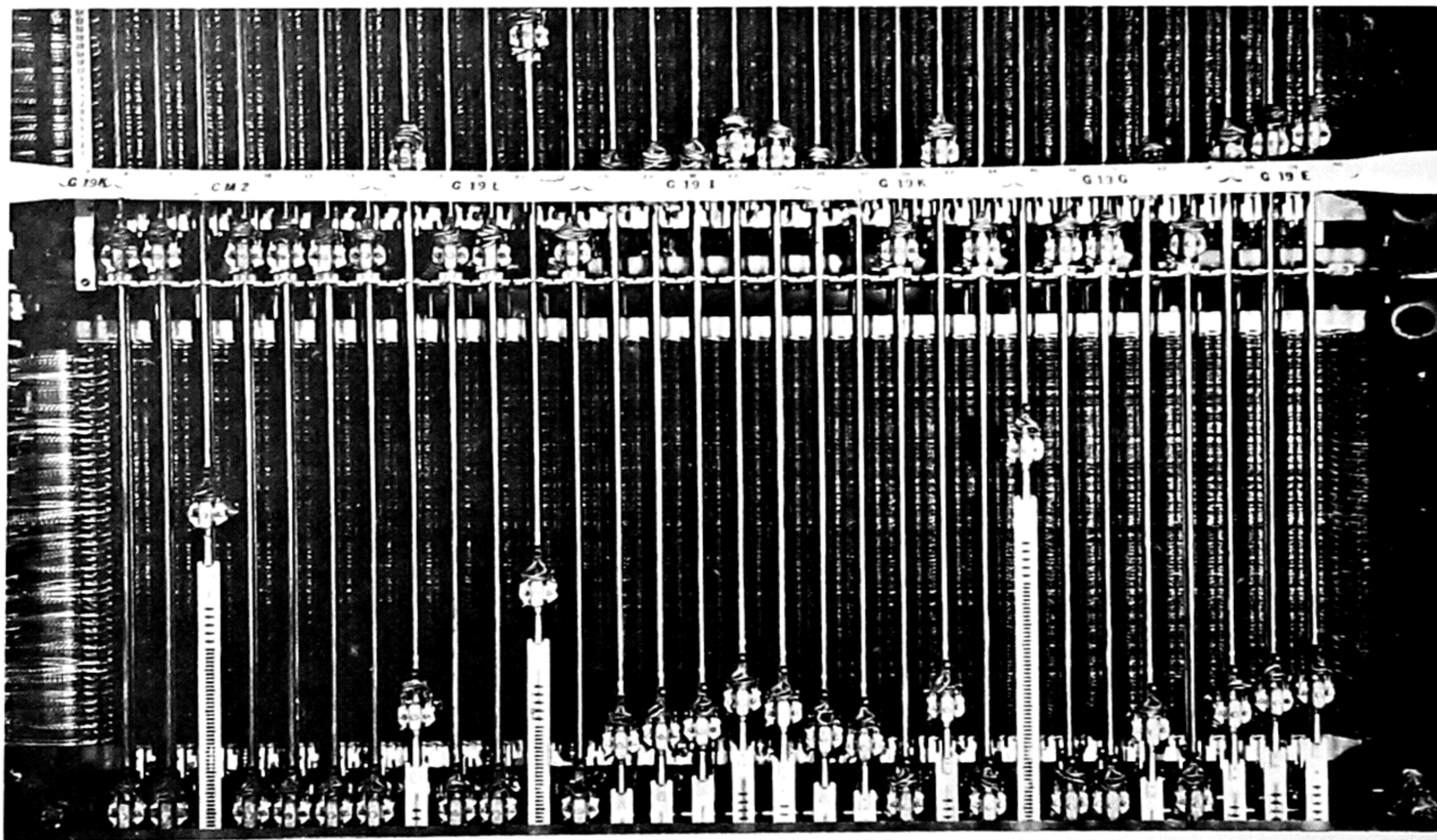
In 1877 an enterprising Boston firm arranged to make telephone connections

among several banks during the day, using the same wires that formed burglar-alarm connections at night. It was not long before private owners began to want to talk with one another too. The first commercial telephone exchange was opened in New Haven, Conn., in 1878.

It accommodated eight lines, interconnected with a series of eight-point switches made of screws and lengths of brass mounted on a board [see illustration on page 697]. There was a push button on each instrument to signal for service but no way of telling the opera-

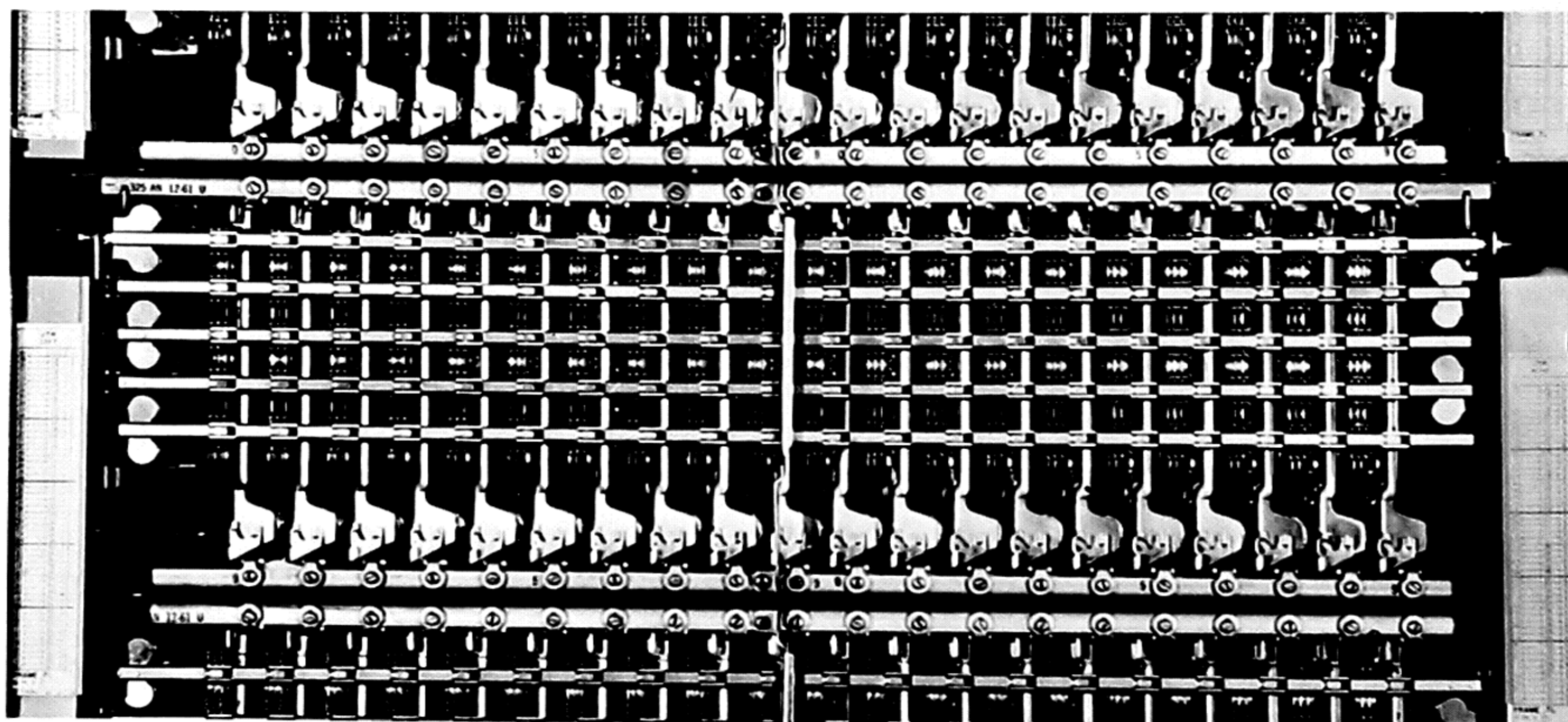


MODERN STROWGER SWITCH consists essentially of a set of contact arms and a semi-circular bank of contacts (bottom half of two switches seen here) and a system of electromagnets and ratchet mechanisms (top half) that control the movement of the contact arms.



THIRTY PANEL SWITCHES lined up side by side constitute the panel bank seen here. The vertical selector rods carry contact

arms over the columns of contacts to the rear. The selector rods in raised positions are in the process of completing calls.



CROSSBAR SWITCH is the most recent development in electro-mechanical telephone switching. A typical crossbar switch is an ar-

ray of 200 contact mechanisms (see top illustration on page 702); these are arranged in 20 vertical columns each having 10 contacts.

tor when a call was completed. He simply listened in from time to time and disconnected the line when he heard no voices.

Soon afterward such crude switches were replaced by boards containing sockets, or "jacks," each connected to a line. The operator had a supply of flexible wires or cords with plugs on both ends. He set up a connection by plugging the ends of a cord into the two jacks associated with the proper lines. Auxiliary circuits between each instrument and the switchboard provided the means to signal for service, ring the telephone being called and notify the operator when the call was completed.

Each switchboard commonly had a capacity of 50 lines, and several boards could be placed side by side to make larger exchanges. One operator could reach over three adjacent boards, thereby serving 150 customers. As the number of customers increased beyond this, more than three boards were needed. Accordingly many calls required the services of two operators and the use of very long cords to set up the necessary connection between different switchboards. It therefore became advantageous to establish permanent connections, called trunks, between nonadjacent boards.

The idea is straightforward, but "trunking" added a new dimension to telephone switching networks. Previ-

ously each position, or jack, on each board led to an individual telephone. There was only one route between any pair of instruments. Now, in addition to jacks representing separate telephones, the operator had a number of more generalized terminals leading to other switchboards. Each of these represented a potential connection to any of a large number of telephones. Moreover, with more than one trunk line connecting each pair of boards, there was a choice of routes. If the operator found that one trunk line was busy, he could try another.

Often, however, a single operator could not keep pace with all the calls coming through his position. Soon a "multiple" switchboard was developed to decrease the load on the operator. Each unit in a multiple board had a complete set of "terminating" connections, one to each telephone line. A second set of "originating" connections to the lines was divided among a number of units [see bottom illustration on following page]. Each operator now handled only a fraction of the incoming traffic, but he could complete a call to any number in the exchange. The basic idea is simple enough, but a great deal of ingenuity went into the design of auxiliary circuits to keep all operators informed as to the status of all lines.

Notice that in referring to operators we have been saying "he." In the early

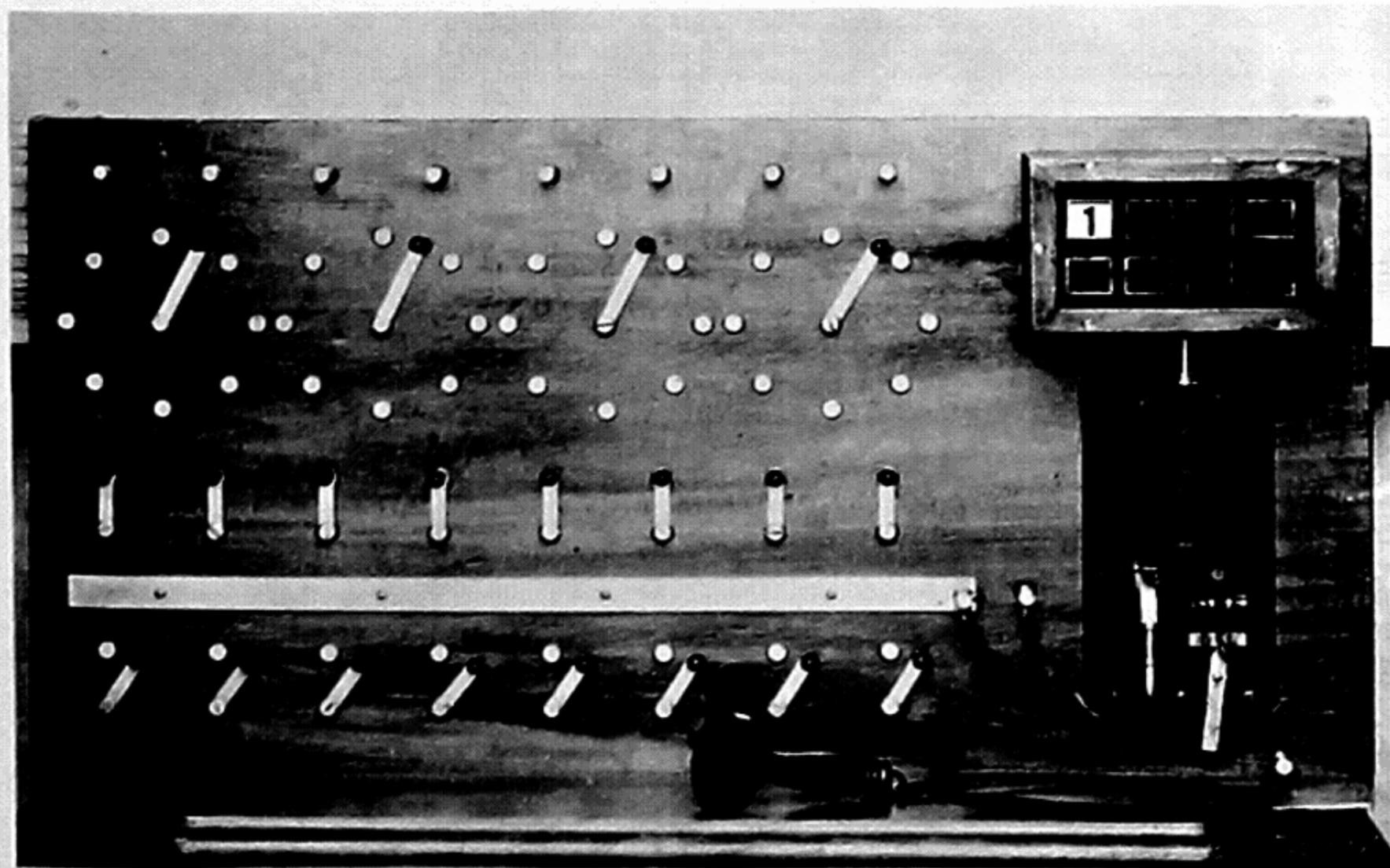
days operators were usually boys or young men. They did not work out well. The youths talked back to customers, shouted, whittled the woodwork on the switchboards and in general created a state of bedlam. The substitution of young women was a complete success. The ladies were quiet, courteous and attentive; they showed more aptitude for the work and remained civil even under considerable provocation.

Once it got started, the telephone network grew at a rapid pace, and engineers began to think of ways to set up connections mechanically. One person who interested himself in the problem was not a telephone man at all but a Kansas City undertaker named Almon B. Strowger. In 1889 Strowger invented a rotary stepping switch that forms the basis for much of the telephone switching equipment in service today.

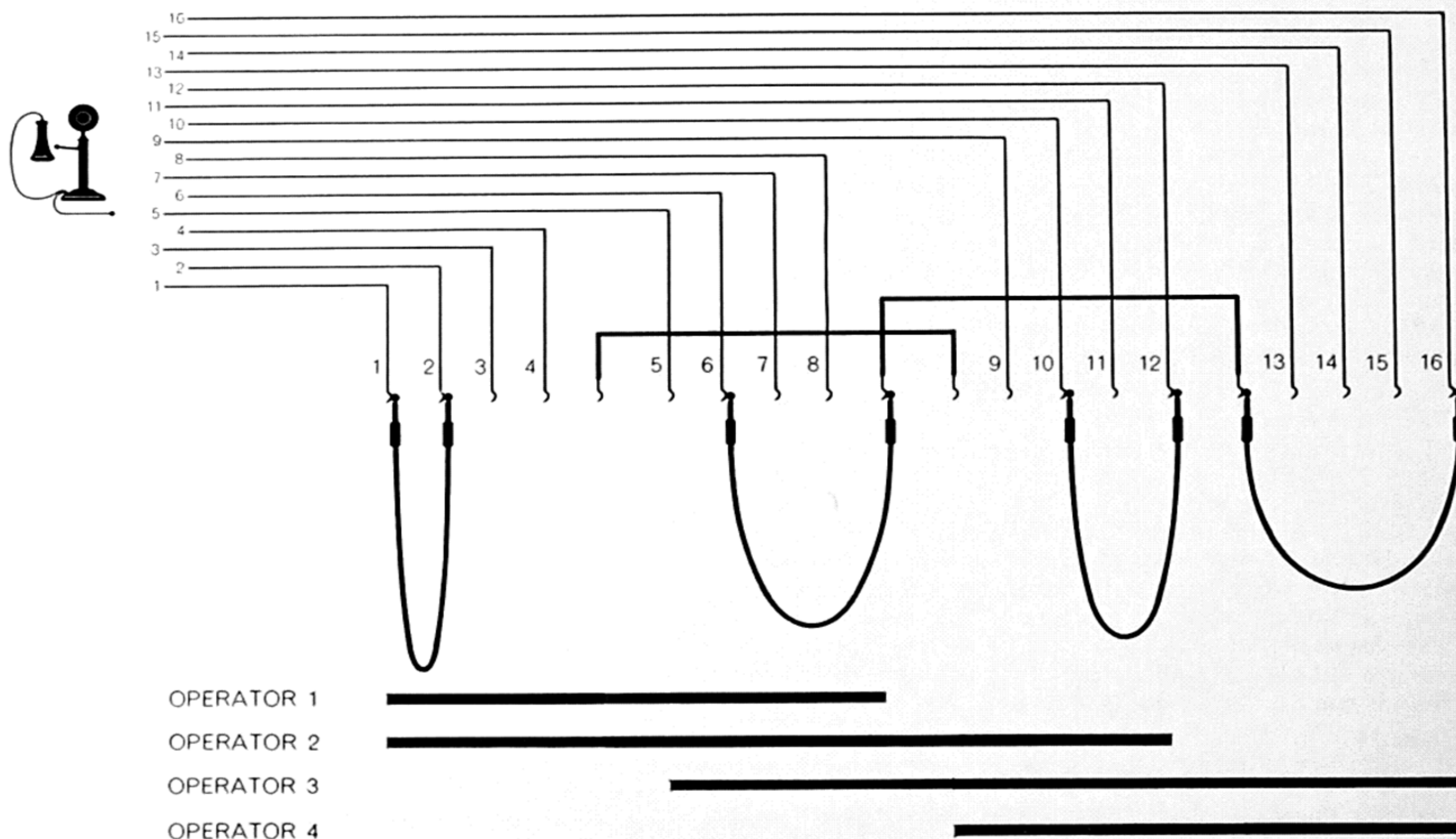
A story, which ought to be true although it may not be, is told of the invention. It seems that many of Strowger's undertaking clients reported getting a wrong number when they telephoned him, and not just any number but the number of a competitor, who proceeded to offer his own services to the callers. By what seemed more than coincidence, the competitor's wife was a telephone operator at the local exchange. Strowger went to the root of the problem: eliminate the operator.

Obviously a man of action, he set about designing a mechanical switch using a collar box and some straight pins for his model. He stuck 10 pins in a row on the inside of the round collar box, with their heads protruding toward the center. Then he pivoted a central arm so that it would make contact with each pin in turn as it rotated. He reasoned that if a group of switches such as this could be made to step from one contact to the next under the control of buttons at the telephone, each customer could set up his own calls. Strowger received a patent on the device in 1891, and it was first installed in La Porte, Ind., the following year.

An improved version of the Strowger switch that is still in use today consists of an array of 10 rows of 10 sets of contacts, each row arranged in an arc of a circle as in the collar-box model. Two simple motions of the contact arm—one vertical, to find a selected row, the other rotary, to find a selected contact in that row—can connect it to any one of 100 lines. The switch is driven by electromagnets and ratchet mechanisms; each time one electromagnet is pulsed, the

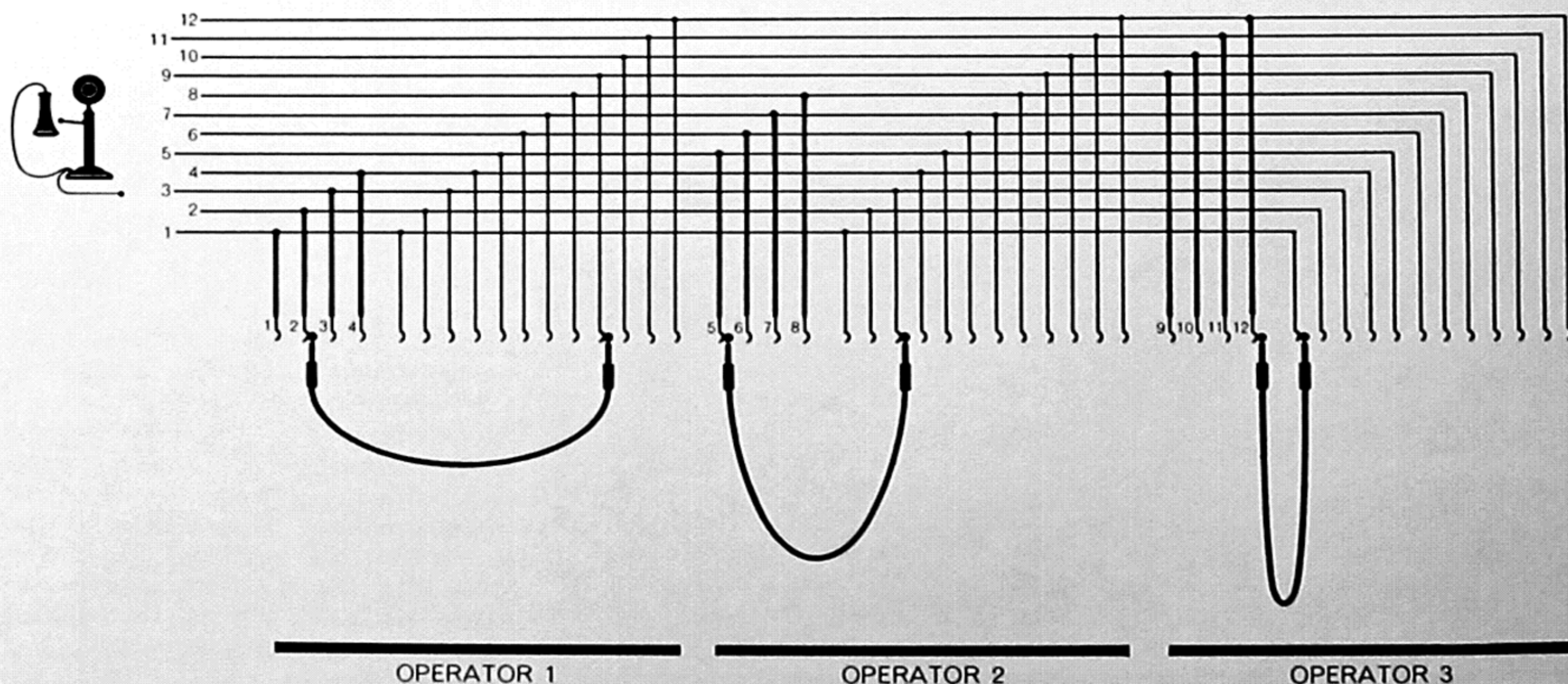


FIRST COMMERCIAL SWITCHBOARD (here represented by a model) was designed for eight telephones. Each of the telephones was equipped with a single push button for signaling the operator, who would then make a connection by setting switches to proper contacts.



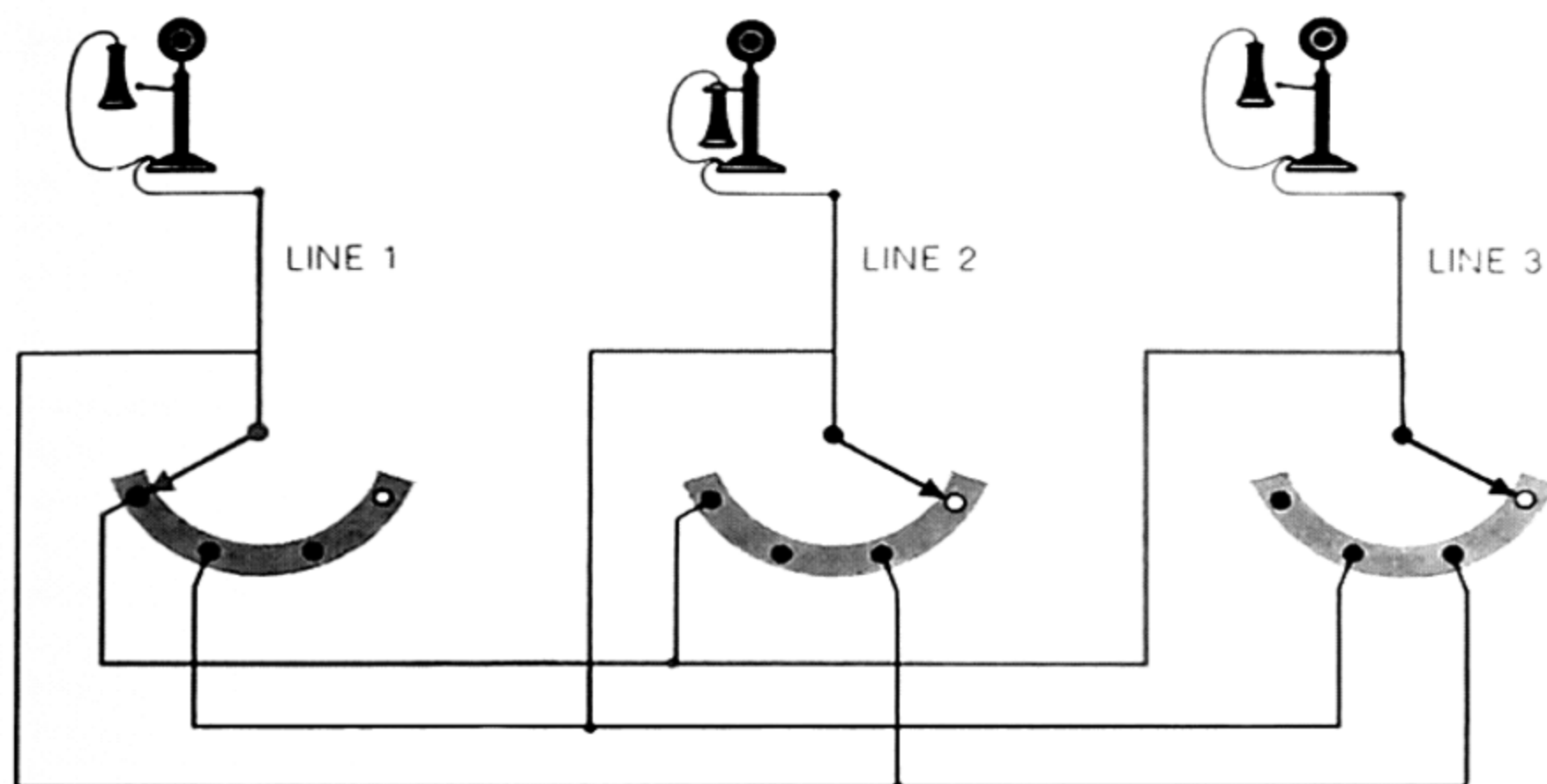
NONMULTIPLE SWITCHBOARD enabled operators to connect two lines in the same unit or in adjacent units. Thus Operator 2 could complete calls for Line 1 through Line 12. But a call on Line

6 for Line 16 required use of the trunk line (*heavy black line*) to Operator 4, who then completed the call. Operator 1 could reach Operator 4 by connecting directly with the same trunk line.



MULTIPLE SWITCHBOARD provides each operator with a limited number of "originating" connections (*vertical heavy black lines*) and a complete set of "terminating" connections. Thus Op-

erator 1 can complete a call on Line 2 for Line 9 by making a direct connection. A nonmultiple switchboard (see illustration at top of page) would have required the use of a trunk line to Operator 3.



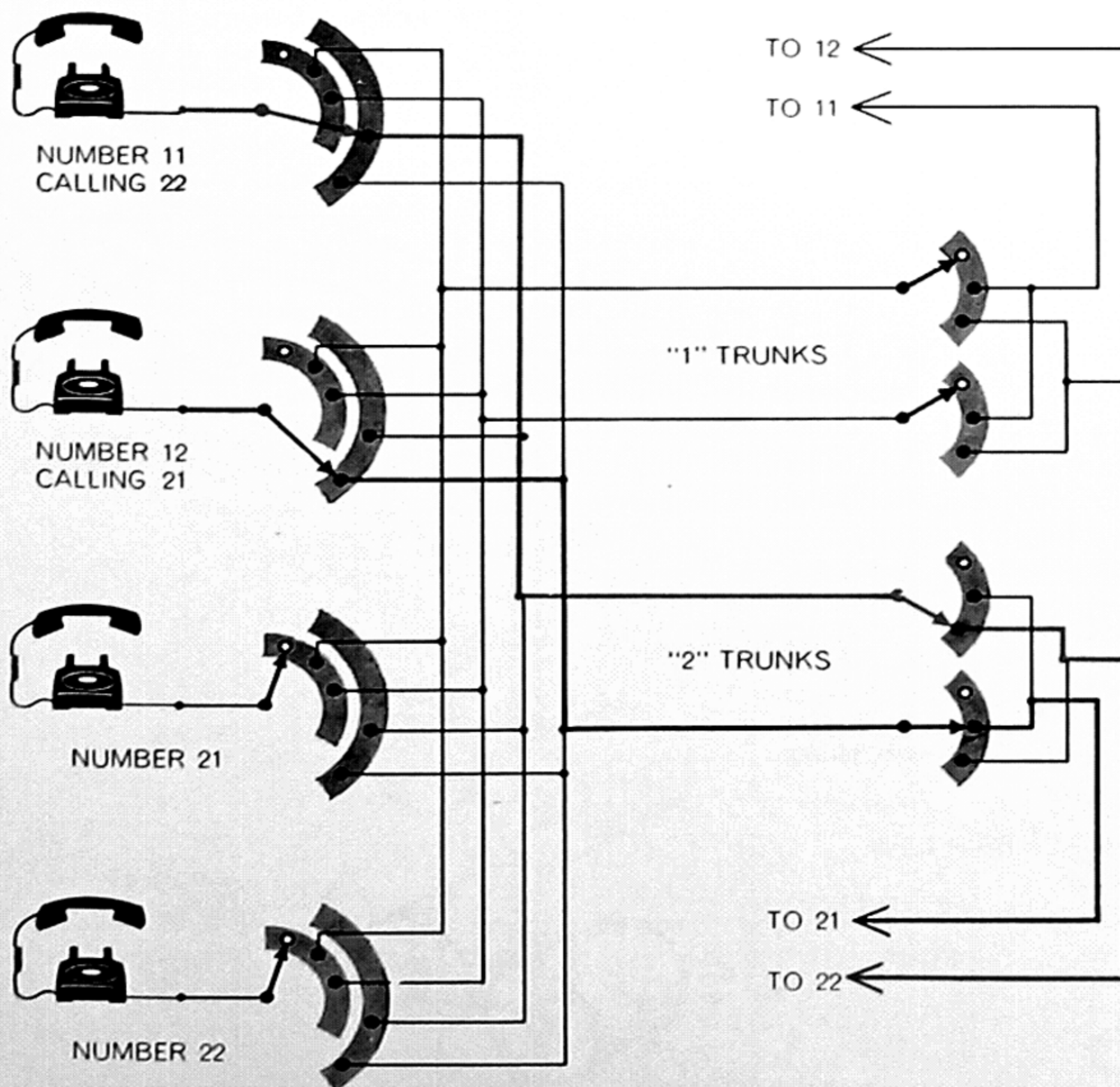
STEP-BY-STEP SWITCHING that set up a connection (*color*) between two phones was based on early Strowger switch (*see illustration at left on following page*). When a subscriber dialed 3 on the first telephone, the Strowger switch associated with that line advanced three steps from the idle position and came to rest on the contact of the third telephone.

arm moves one row vertically, each time a second electromagnet is pulsed, the arm rotates one position horizontally. A third electromagnet releases the arm, letting it return to its starting position by spring action and gravity.

In the first systems employing the Strowger switch the telephones had no dials but small boxes with three push buttons, one for each electromagnet. Each instrument was connected to the arm of its own switch. Each contact in that switch was connected to one of 100 telephones. Repeated operation of the push buttons caused the arm to step along from row to row and from contact to contact so that the customer could set up a direct connection to any of 99 others. This arrangement, technically known as a step-by-step system, was popularly known as the "girl-less, cussless telephone."

FIRST DIGIT SWITCHES

SECOND DIGIT SWITCHES



"HUNTING" available routes for two calls is illustrated. Dialing the first digit of 22 on Number 11 raises the contact arm to the contact bank leading to the "2" trunks; the arm hunts out the first nonbusy contact. Dialing the second digit then connects the call to Number 22. When 21 is dialed on Number 12, the first trunk line is busy and the arm hunts out the next nonbusy trunk. The dialing of the second digit connects the call to Number 21.

In a self-contained network of 100 telephones step-by-step switching allowed each customer to do exactly what an operator would have done in setting up a call. How could the principle be applied to a system involving trunks? How could the customer know which of various routes to a given destination were available and which were busy? The answer to these questions constituted a giant forward stride: The switching network itself was designed to participate in the selection of a route.

Under the new system each telephone originating a call was connected to the pivoting arm of a switch. The successive banks of contacts of that first switch, known as a selector, no longer led to individual telephones but to groups of trunks. In other words, the lowest 10 contacts were connected to 10 trunks running to one group of switches, the next row provided 10 routes to a second group, and so on. By dialing one digit (dials had quickly replaced push buttons) the customer selected a particular row of contacts, as before. When the arm reached the row, it proceeded to "hunt" across it automatically: it stepped from one contact to the next until it found one that was not busy, then stopped there. This action was made possible by coupling two similar contact arms on the hunting switch so that they moved together. Each of the arms had an associated bank of contacts. The first bank was used for the desired connection. Each contact in the duplicate bank carried a signal indicating whether its corresponding member in the original bank was busy or idle. If the signal showed busy, a pulse was delivered automatical-

ly to the second electromagnet mentioned earlier, thus causing the switch to step horizontally along the row. When a contact carrying the idle signal was reached, no further pulse was delivered and the switch rested there.

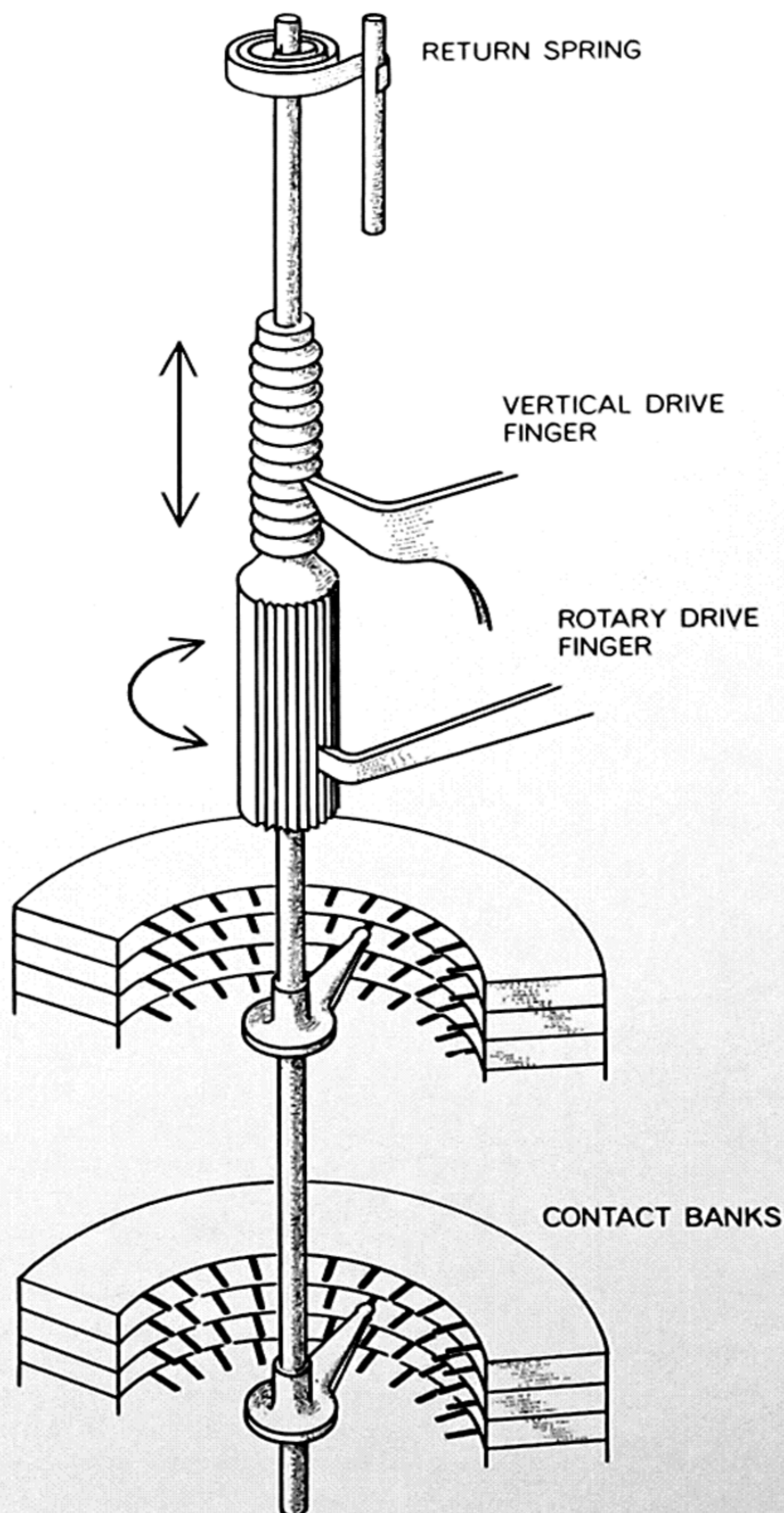
The customer now had a trunk line to the group of switches serving the number he was calling. At the end of the trunk was another switch leading to the individual telephones. Further dialing pulses operated that switch—both vertically and horizontally—to make the de-

sired connection. The entire procedure was still largely under the control of the person dialing, but hunting increased the efficiency of operation and made possible larger systems of 1,000 lines or more.

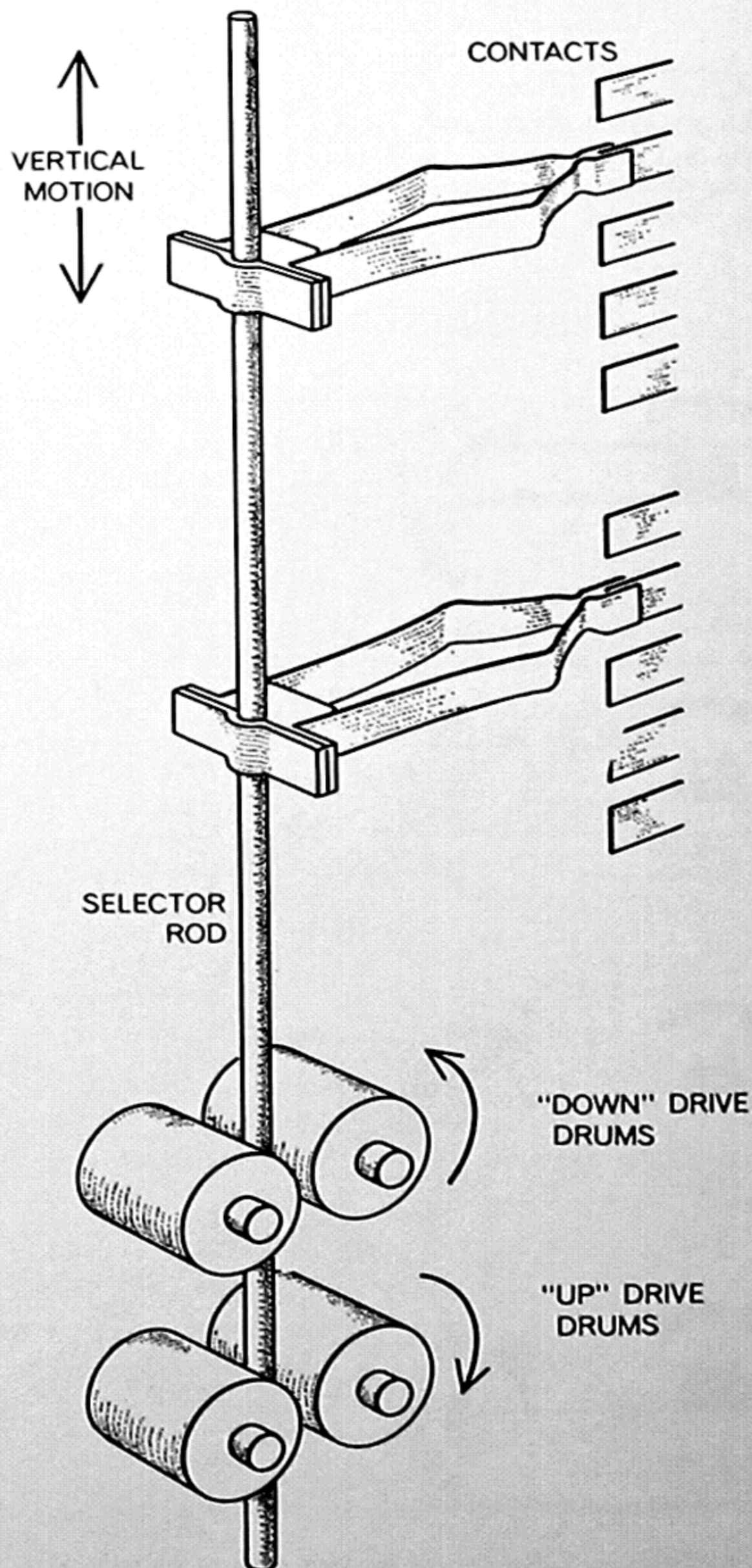
The hunting concept was also applied to reduce the number of switches required in a system. Formerly each telephone was connected directly to an individual selector switch. A new arrangement introduced a "line finder" switch that hunted over a group of lines, finding

one that was requesting service. Through the line-finder switch that line would then be connected to one of a group of selector switches. In this way switches no longer had to be supplied on a per-line basis; they were provided in smaller quantities appropriate to the traffic.

In addition to improving the efficiency of switches and decreasing their number, engineers extended the trunking idea to more and more complex branching networks with several stages of in-



ORIGINAL STROWGER SWITCH made possible systems of 100 telephones each. Vertical and rotary motion of the contact arms in step-by-step fashion connected one telephone with any of 99 others.



PANEL SWITCH has several flat banks of contacts with as many as 500 terminals in a column. Any terminal can be reached by contact fingers of a selector rod controlled by motor-driven drums.

intermediate selector switches. As the branching increased, the number of separate pulsing signals necessary to operate all the switches grew. Eventually numbers were standardized with three-digit codes (two of which were commonly labeled by letters) for central offices and four-digit numbers providing access to as many as 10,000 telephones in each office.

These advances in technique solved the major problems and made large switching systems serving many thousands of lines economically feasible. Step-by-step systems employing the Strowger switch are still used in many central offices, usually the smaller ones. The arrangement allows little flexibility in the assigning of telephone numbers, because the numbers must correspond to the location of particular terminals on the switches. In addition, the switches in the early step-by-step systems we have described are directly controlled by the dialer; in a sense he operates them by hand. As a result the entire assemblage is geared to the manual speed of the person dialing a number. Yet even these fairly rudimentary switches can work

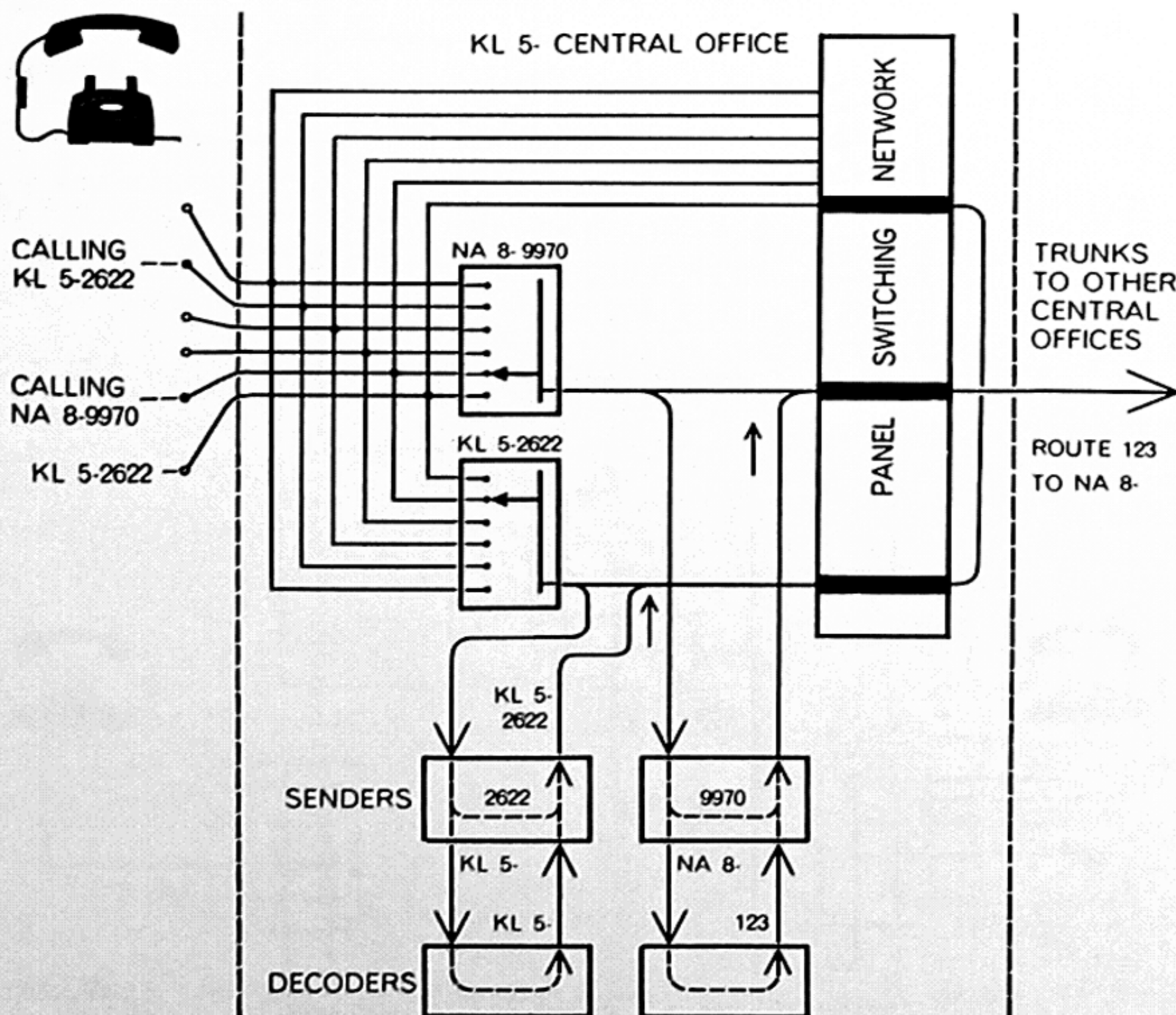
a good deal faster than that, and more advanced devices operate literally millions of times faster. The amount of equipment tied up in completing a single call under direct control is inherently capable of handling hundreds of calls in the same length of time.

Furthermore, Strowger switches provided access to one of only 100 output terminals. Increasing the number of possible connections per switch would reduce the number of stages through which a call had to be routed and would also allow hunting over larger trunk groups. The next switch to be developed, the panel switch, has 500 contacts arranged in a single vertical column and contact fingers that move up and down along them. Unlike the moving arms in Strowger switches, the fingers are not driven a step at a time by ratchets but move smoothly on motor-driven shafts controlled by clutches [see illustration at right on preceding page]. Therefore these switches are not controlled directly by dial pulses, as are the Strowger switches. To control the panel switches and to speed up system operation, a new element, called a "sender," is intro-

duced. The sender records the incoming dialed number and then uses it to control the panel switches at a much higher speed.

Used with the sender is a decoder. The function of this decoder is to convert the decimal numbers received from the dial to the nondecimal system required by the switches. In addition the decoder provides flexibility in rearranging trunk connections. Panel switches, in combination with sender-decoders, became standard equipment for larger installations. Although they are no longer being installed, many are still in service in New York, Chicago and other big cities.

Panel switches are much more efficient than their predecessors, but they too have drawbacks, mostly connected with the long and frequent excursions of their contact arms. The excursions take time and wear out the clutches and other moving parts. Continual "fingering" of contacts causes electrical noise, particularly since the contacts tend to collect dust and dirt. For this reason the panel devices have now been displaced by the "crossbar" switch, which represents the last word in electromechanical switching.



PANEL SWITCHING SYSTEM described in the text removes the switching process from the dialer's direct control by interposing "senders" and "decoders" between the dial and switching systems. The call's route is set up automatically and thus much more quickly.

The crossbar consists of a grid or matrix of 10 horizontal and 10 vertical wires, with one group slightly in front of the other so that they do not touch. At each intersection is a movable contact. As many as 10 of them can be closed at one time to connect any of the horizontal wires to any of the vertical wires, providing 10 separate paths through the grid. The contacts are ingenious combinations of springs and levers that open or close in response to the combined rotary motion of one of five horizontal bars and linear motion of one of 10 vertical bars. Only 20 electromagnets are required to control the 100 contacts: two magnets for each rotating bar (one for each direction of rotation) and one magnet for each bar moving linearly. The action of these switches involves small, rapid mechanical motions and is adapted to straightforward selection schemes for closing the contacts. Crossbar systems offer advantages beyond those inherent in the switch itself. The trend toward separation of switching and control functions, which began with the introduction of senders and decoders, has been carried all the way. In modern crossbar systems the network switches have nothing to do with hunting or path selection. Control is centralized in a group of circuits called

markers because they mark a path through the network of crossbar switches. The arrangement is very flexible and makes more decisions in less time than earlier systems could. In addition it offers many new services.

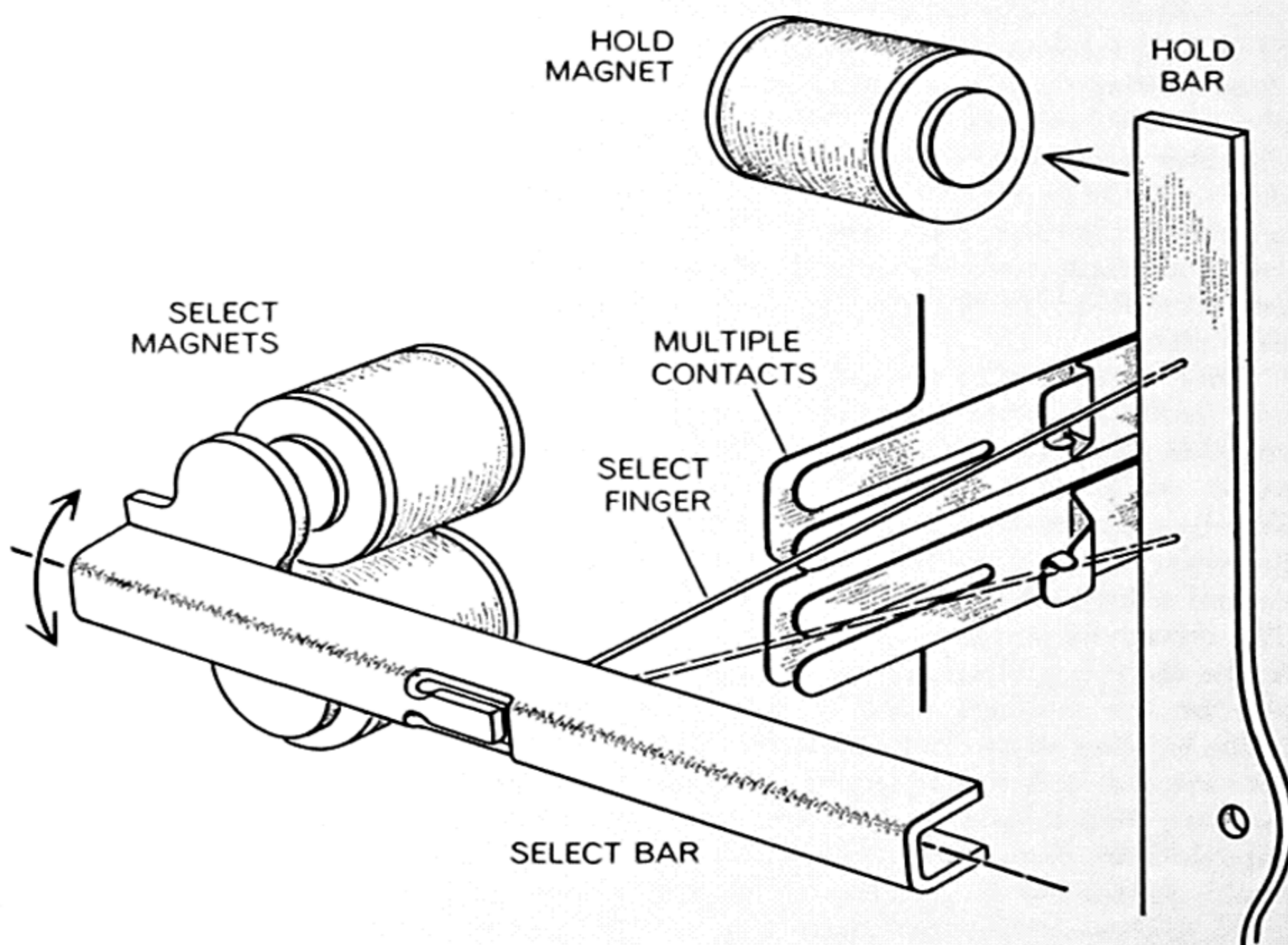
One of these is "alternate routing." If the marker finds that all the trunks going directly to a particular central office are busy, it can select an alternate route via another nearby office, which in turn picks out an idle trunk to the desired destination. The system also makes a second trial if an attempt to set up a call is thwarted. When one part of the network is busy or has failed, it will try another. A third valuable feature is the ability of the system to detect failures within its own equipment and report them on punched cards. This greatly simplifies the task of troubleshooting the complex circuits.

In the crossbar system electromechanical switching has gone about as far as it can go. The major advances of the future lie in electronic circuits, which appear to offer even greater flexibility and still higher speed. They should make possible a reduction in the size and cost of switching systems in addition to providing more versatile service.

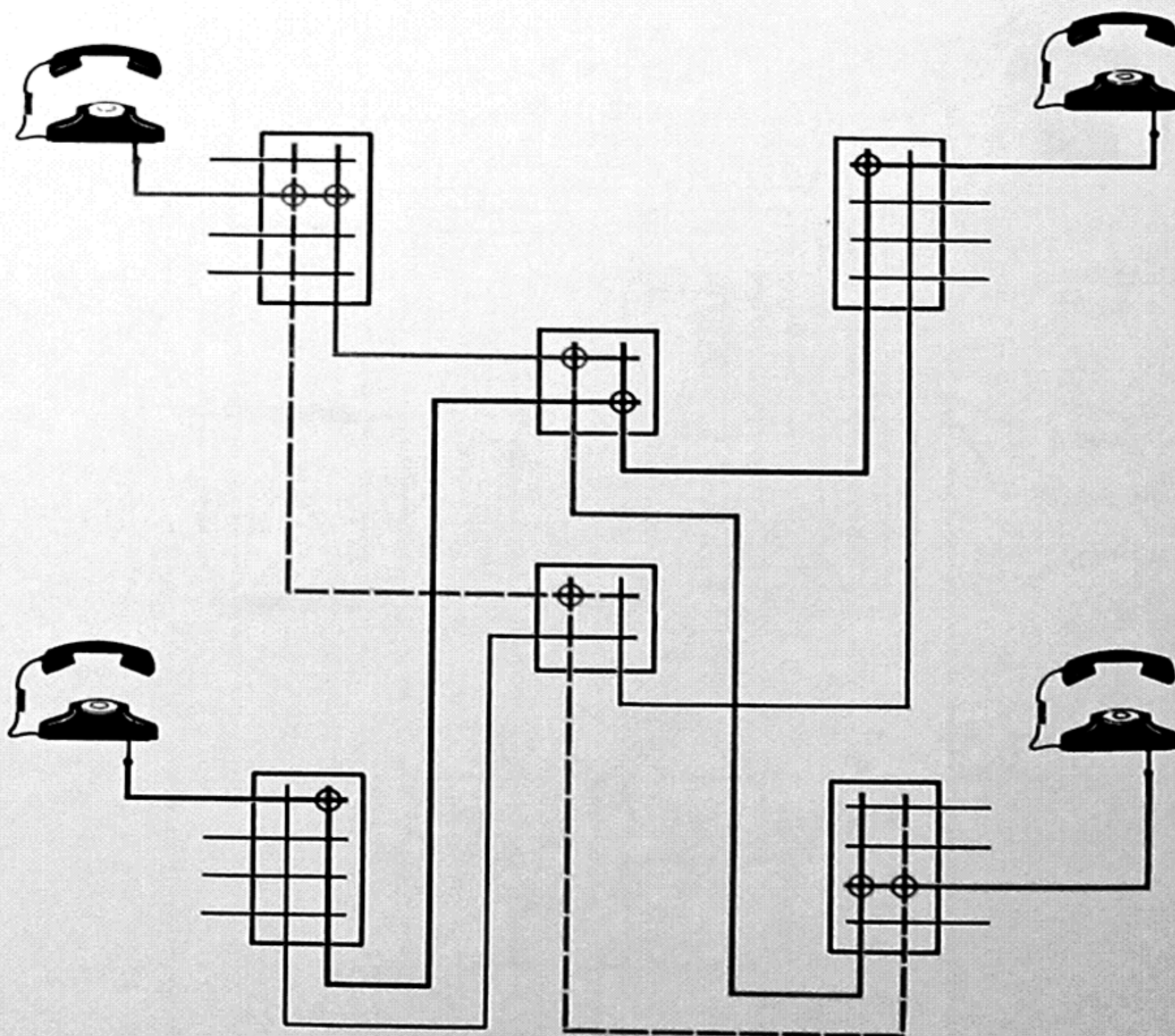
The basic unit for electronic switching is the transistor. It is a simple on-off device: a signal applied to a control electrode opens or closes the circuit between two other electrodes. A transistor is therefore equivalent to a mechanical switch with only a single contact. To enable transistors to compete with today's fast-acting multiple contact switch it is necessary to exploit to the full their high speed. Mechanical relays switch in thousandths of a second, transistors in millionths of a second (microseconds) or less. By capitalizing on the 1,000-to-1 advantage it is possible to build a single electronic unit that replaces several markers and other items of control equipment in crossbar systems.

Electronic switching has also clarified the role played by logic and memory. Obviously the earliest control systems could store information and carry out logical operations, but the fact that they did so was somewhat obscured by the mechanical detail. In electronic systems these functions stand out clearly.

Controlling any switching system, no matter how complex, in reality involves nothing more than the making of a large number of simple logical decisions. For example, suppose that the control is informed by the appropriate monitoring circuit that a particular telephone is off



CROSSBAR SWITCH CONTACT MECHANISM operates by the combined motion of a "select bar" and a "hold bar." The motion of the hold bar closes the contact only when the "select finger" has been placed in either of two positions by the movement of the select bar.



CROSSBAR SWITCHING SYSTEM uses space division to route telephone calls. Two typical paths, in color and in black, are spatially separated, although they appear to cross at some points. The broken line in color represents an alternative path for one of the calls.

the hook. What action should be taken? In order to decide, the control circuit first determines whether the telephone is in a talking connection or is connected to a dial tone. In either case "off hook" is a proper line condition and no action is required. If neither situation exists, the line has just gone off hook and is requesting service. The control device then arranges to supply dial tone to the telephone. Similarly, the decisions of the control in response to any other condition can be spelled out in terms of the present state of the call and new information received from hook and dial.

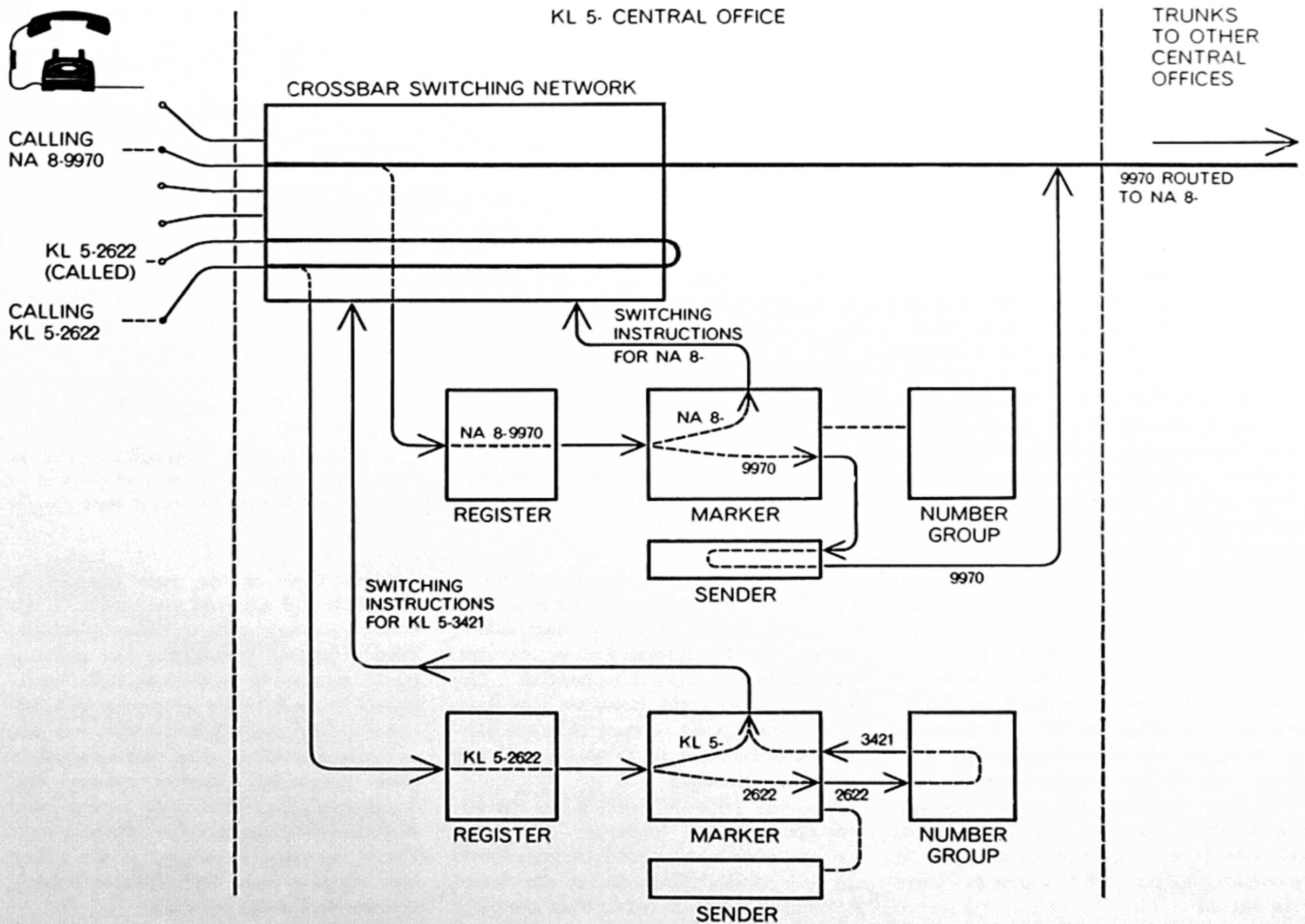
There are two approaches to the problem of building logic into control systems. One might be called wired logic. Each logical operation is clearly defined

and a circuit is designed to do the specific job. As an illustration of simple wired logic circuits, consider a light controlled by a pair of switches. If the switches are wired in parallel, the light goes on when either one or the other is snapped. The circuit embodies a logical "or" function. If the switches are wired in series, the operation of the first switch *and* the second switch is required to turn on the light—a logical "and" function. Elementary "or" and "and" units can be combined in large circuits capable of deciding complicated logical questions.

Wired logic circuits make virtually all decisions in electromechanical switching systems, and they do an excellent job. They can be built to take advantage of

the many contacts that are available in relay and other mechanical switches. But each circuit is tailor-made for a single operation; to make a change usually requires extensive reworking and rewiring. Furthermore, if electronic logic elements are used, the circuits often become quite expensive.

The second type of control, called a stored program, consists essentially of a special-purpose computer with two memories. One is a semipermanent memory in which are stored the processing routines for setting up calls. The procedures are "written" not in copper wire and solder, as in wired logic, but in the form of magnetic patterns on plug-in cards. Changing one of the logical procedures is then accomplished simply by



CROSSBAR CENTRAL OFFICE employs registers, markers and senders. For a call within the central office, the register receives the originating dial pulses (*KL 5-2622*), stores the information and passes it on to the marker. The marker sends the last four digits to the number group (an electronic "telephone directory"),

getting back a code number; it uses the new information (*KL 5-3421*) to select a path through the switching network and then closes the proper switches to complete the call. For a call outside the central office (*NA 8-9970*) the sender routes the last four digits to the appropriate office along a path selected by the marker.

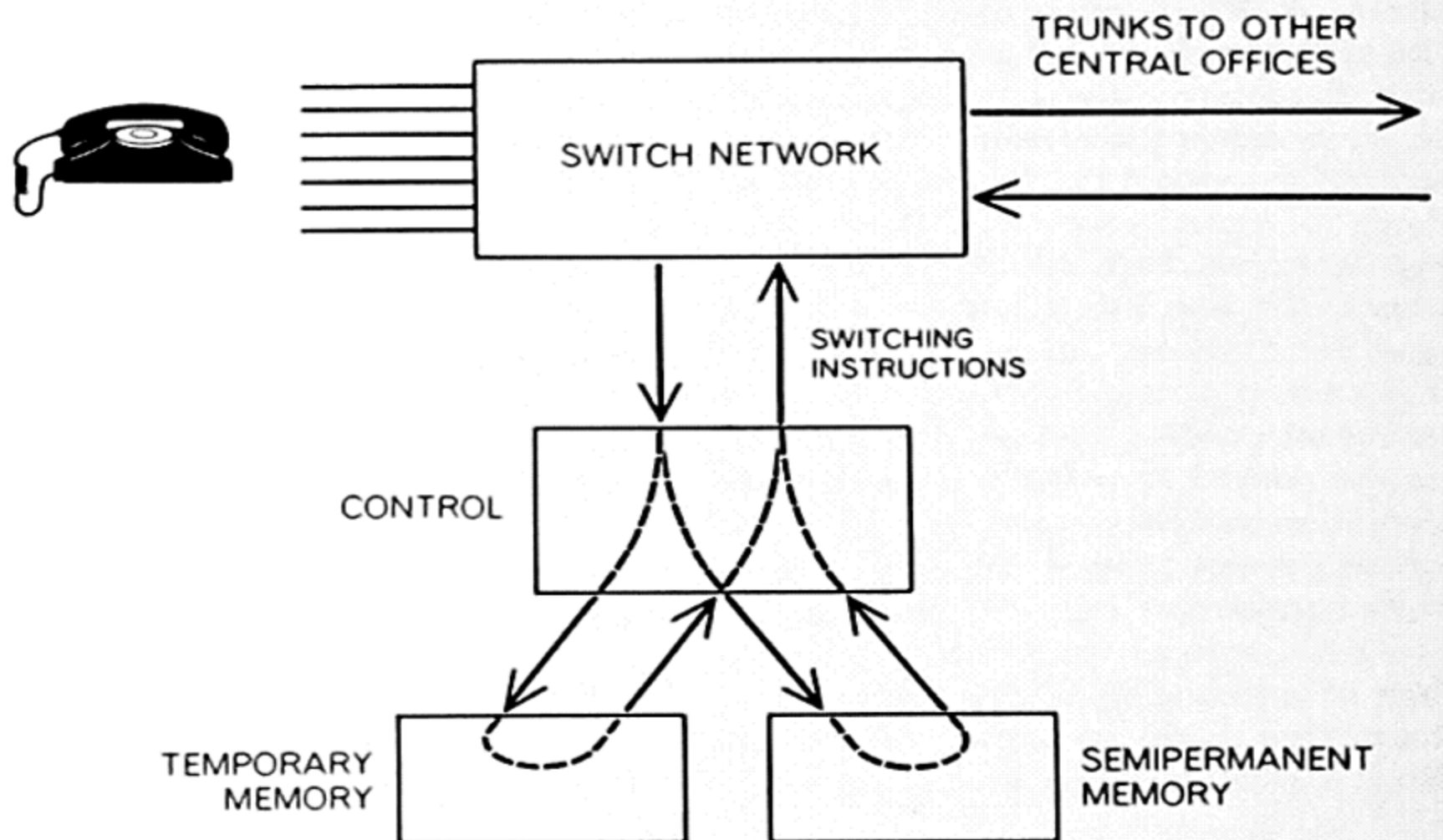
substituting a new card. The second memory stores short-term information such as the present state of various calls. It may be thought of as an electronic scratch-pad.

As our earlier examples have shown, the processing of a telephone call can be analyzed into a number of discrete situations, each calling for a logical decision as to the next step. A telephone off hook constitutes one situation; two telephones connected for talking, a second situation; one instrument put back on hook, a third, and so on. Each such situation—and there are hundreds—may be assigned a number. The semi-permanent plug-in card memory lists each situation number, together with the different actions that may be required, depending on the additional information coming in from the customers by way of hook and dial.

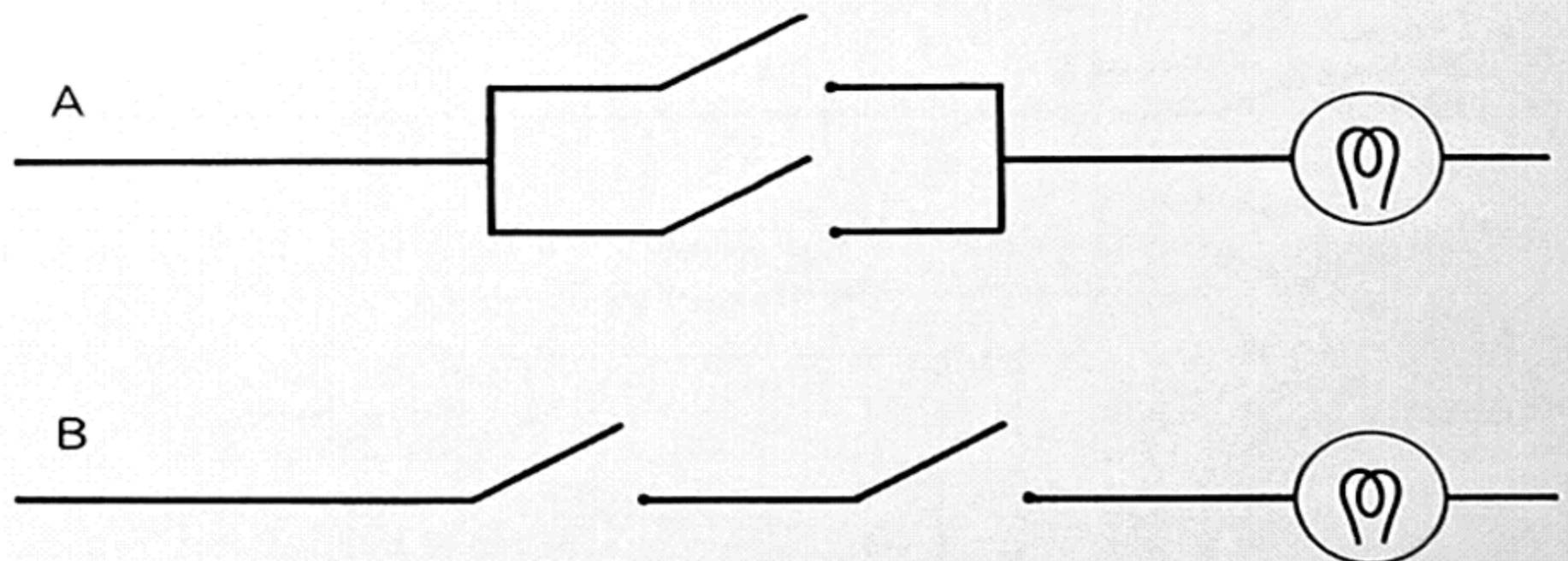
This information is recorded in the scratch-pad memory as it is received. Also listed on the scratch-pad are the states of the switches in the network. The content of the short-term memory continually changes to reflect the progress of the calls the system is working on.

In oversimplified terms, the processing of a call goes something like this: A customer takes his phone off hook to start the call. Assume this is labeled Situation No. 1. Space is now assigned on the scratch-pad for the call, and the calling telephone number and Situation No. 1 are written in this space. The scratch-pad memory is electronically interrogated at frequent intervals. At the next interrogation Situation No. 1 is read out and checked with the program stored in the semipermanent memory. The instructions found there are to determine whether the telephone is answering a call or originating one. The information can be obtained by referring to the scratch-pad memory. In the present example the scratch-pad will show that the off-hook telephone is not connected to another telephone or to dial tone; in other words, that this is a call origination. Accordingly the program will instruct the network to connect the telephone to an originating register that supplies dial tone. After this has been done the scratch-pad is revised to list the call as being in Situation No. 2.

As each dialed digit is received, it is recorded in the scratch-pad memory. The digits can then be read out to enable the stored program to choose a path through the network. As each call is acted on, the scratch-pad notations are



ELECTRONIC SWITCHING SYSTEM is simplified here. The temporary memory stores dial pulses. Using this and processing information stored in the semipermanent memory, the control circuit selects a path through the switch network and closes the proper switches.



SIMPLE LOGIC CIRCUITS are illustrated by a light controlled by switches wired in parallel (A) and in series (B). A represents a logical "or" function (closing one switch or the other turns the light on); B, an "and" function (the first and second switch must close).

changed to read Situation No. 3, No. 4, No. 5 and so on as appropriate. The scratch-pad has a space for each call in process and the information in the various spaces is examined sequentially. The sequence is so rapid, however, that from a customer's point of view calls are handled side by side, so to speak, and not one after the other.

A stored program control has the advantage of great flexibility. Only the size of the semipermanent memory limits the new routines that can be introduced on plug-in cards. Services that were not even thought of when the system was designed can easily be installed as the need becomes clear.

Recently an electronic telephone switching system was tried out in Morris, Ill. It used a stored-program electronic

control. One of the new services it demonstrated allowed customers to re-route incoming calls to other numbers. People visiting friends for the evening could arrange to have their calls transferred to their host's telephone by dialing a special code. The transfer was just as easily discontinued by dialing another code. A second feature replaced the usual seven-digit dial code with a two-digit code for numbers frequently called. These are merely samples of the many new services that electronic switching systems will make possible.

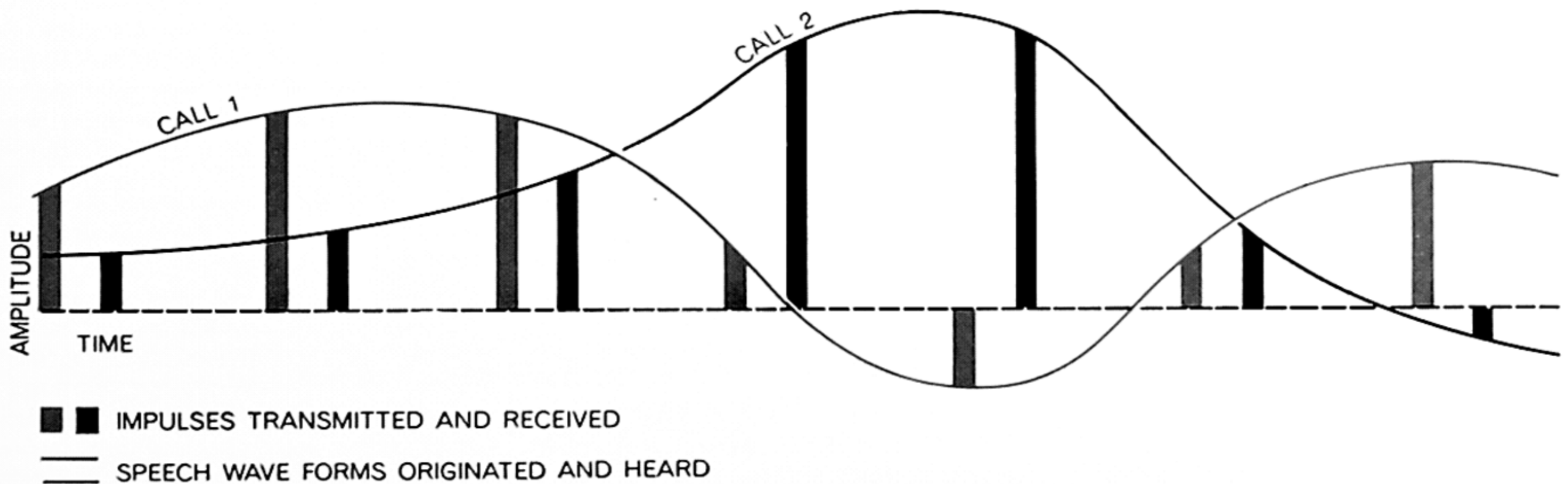
The Morris experiment demonstrated beyond question the technical feasibility as well as the great versatility of electronic systems. There is, however, still a great deal of activity in the field of telephone switching. One area in which

there is intensive study has to do with a different type of switching network altogether—one that only the great speed of the transistor has made possible.

All the switching networks we have discussed so far have had one property in common: the pathways that exist at any one time are physically separate. At no place do two calls travel the same route simultaneously. For that reason such systems are said to have space-division networks. The new approach, now being studied at the Bell

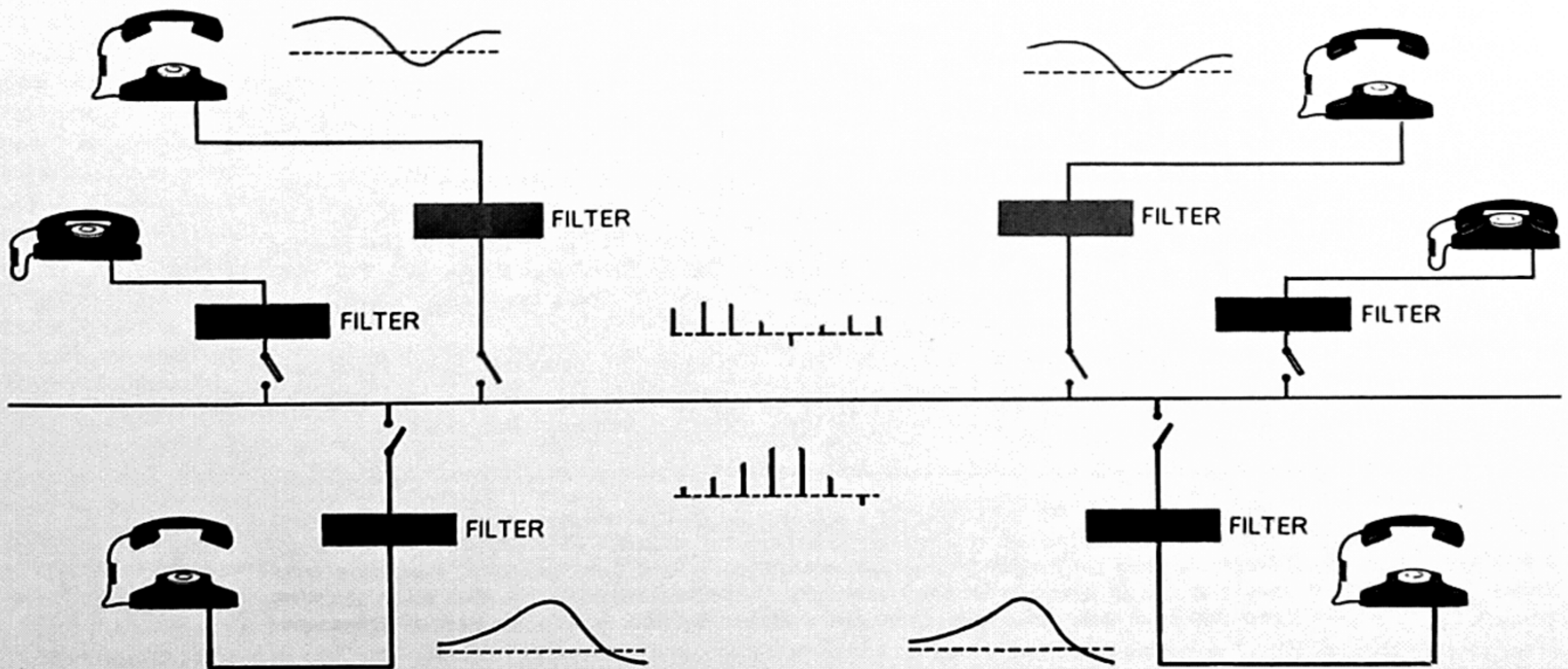
Telephone Laboratories and elsewhere, is known as time division. Calls travel over the same path at the same time, separated from one another in time. This is how the scheme works: When a connection is established between a pair of telephones via the common path, the two switches leading into and out of this path do not stay closed all the time. Instead they open and close continually at a rapid rate, being closed only a small fraction of the time, typically two microseconds in every 100. Therefore they send through the common path not the

complete electrical wave coming from the transmitting telephone but a series of short pulses that are samples of that wave [see upper illustration below]. Each pair of switches, for each pair of connected telephones, works the same way, but their times of closing are displaced with respect to one another so that the sampling is done in rotation and signals from all the calls are separated in time. The approximate capacity of the common path equals the time between closures of any pair of switches divided by the time for which each pair



SAMPLES OF SOUND (impulses transmitted and received) are transmitted or received in a time-division telephone system. Al-

though only two telephone calls have been represented, a common path is theoretically capable of carrying many simultaneous calls.



TIME-DIVISION SWITCHING NETWORK works on the repetitive and simultaneous opening and closing of selected pairs of switches. When the switches are open, a filter stores the energy

from the sending telephone and, when they are closed, transmits this energy to the receiving filter. This filter converts the pulses into a smoothed-out wave and feeds it into the receiving telephone.

remains closed. In the present example 100 would be divided by two, giving 50 simultaneous calls.

Although the switches for each connection are closed perhaps only 2 per cent of the time, the signal reaching the receiving phone is complete and essentially undistorted. In fact, theoretically it is transmitted without any loss of information whatever. The reconstitution of the complete wave, surprising as it may seem, is analogous to the effect we see whenever we go to a movie. The picture is a series of still "samples" of motion; the eye smooths over the intervals between stills and restores the motion.

In time-division telephone switching networks the smoothing action is accomplished by electrical filters. There is one between each telephone and its switch. While the sending switch is open, the

filter stores the energy coming from the telephone. When the two switches are closed, this energy is transmitted via a pulse to the receiving filter, which is identical with the sending device and caused to be in resonance with it. The "resonant transfer" provides theoretically lossless transmission. (In practice there is a small loss.) The receiving filter in turn converts the incoming pulses to a smoothed-out wave, which it feeds into the receiving telephone.

The system can be designed to meet any required fidelity; that is, the final wave can be made as nearly identical with the original as is desired. To obtain adequate fidelity the frequency at which the switches close—the sampling rate—must be somewhat more than twice the highest frequency transmitted. If the highest transmitted frequency is 3,500 cycles per second, a usual figure in

telephone work, the sampling frequency should be something more than 7,000 cycles per second. In practice it is usually set at 8,000 cycles or more.

Time-division switching offers the main advantage of requiring fewer switches than space-division schemes for systems of equal capacity. The advantage is partially offset by the need for two filters, as well as some other additional components, for each pair of sampling switches.

There is no question that both time-division and space-division electronic switching networks are technically feasible. Both schemes must be considered strong contenders, and each may find its own area of usefulness. Although the cost and reliability of electronic switching systems are still being investigated, there is little doubt that the great benefits they offer will soon be realized.

The Authors

H. S. FEDER and A. E. SPENCER are respectively a member of the technical staff of the Electronic Switching Division and a supervisor in the Military Communications Systems Engineering Center at the Bell Telephone Laboratories. Feder took his B.S. in electrical engineering at the College of the City of New York in 1944 and joined the Bell Laboratories in 1946, having worked in the meantime for the Western Electric Company and at the Los Alamos Scientific Laboratory. He received an M.S. in electrical engineering from Rutgers University in 1960 and is currently doing research on the "time division" telephone system discussed in the present article. Spencer obtained a B.S. in electrical engineering from Drexel Institute of Technology in 1951. He went to the

Bell Laboratories the same year, where he is engaged in the design of communications systems.

Bibliography

- BEGINNINGS OF TELEPHONY. Frederick L. Rhodes. Harper & Brothers, 1929.
- COMMON CONTROL TELEPHONE SWITCHING SYSTEMS. Oscar Myers in *The Bell System Technical Journal*, Vol. 31, No. 6, pages 1086–1120; November, 1952.
- EFFICIENCY AND RECIPROCITY IN PULSE-AMPLITUDE MODULATION. K. W. Cattermole in *Proceedings of the Institution of Electrical Engineers*, Vol. 105, Part B, pages 449–462; September, 1958.
- TELEPHONE THEORY AND PRACTICE. Kempster B. Miller. McGraw-Hill Book Company, Inc., 1933.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

THE SPARK CHAMBER

by Gerard K. O'Neill

It is a new device to make visible the tracks of subatomic particles. Basically it consists of a series of charged plates. When a particle passes through the plates, its path is traced by sparks between them.

The present understanding, imperfect but growing, of the fundamental nature of matter has come largely from observation of the elementary particles. The protons, neutrons, electrons, mesons and other particles reveal the most when they can be studied one at a time or when only two or three of them interact. When larger numbers are present, the sheer mathematical complexity of their interaction hides the fundamental simplicities. For this reason the efforts of many experimental physicists over several decades have gone into the development of sensitive methods for detecting single particles.

There is no single best design for a particle detector. To obtain certain characteristics it is usually necessary to sacrifice others, and the choice depends on the nature of the experimental "events" one wishes to observe. Physicists working with the large particle-accelerating machines have increasingly been concerned with extremely rare events, epitomized by the recent discovery at the Brookhaven National Laboratory that there are two kinds of neutrino rather than one. To obtain the evidence for this discovery the 30-billion-electron-volt proton accelerator at Brookhaven was operated for six months. Over this period the number of recorded events caused by neutrinos averaged fewer than one every three days. The particle detector used in the experiment is of an entirely new type: it is called a spark chamber. Before explaining its operation I shall describe the general nature of the particle-detection problem.

The problem is far from easy, because an elementary particle can pass freely through many atoms of any substance without leaving a trace. Even at present there is no practical device that can detect electrically neutral particles without destroying or deflecting

them. Charged particles, however, exert a strong electrostatic force on the electrons of the atoms through which they pass. Usually the electrostatic force between the negative electron and the positive nucleus is enough to keep the electrons from breaking free, but occasionally—roughly once in every 1,000 atoms through which a charged particle passes—an electron is jolted loose. In air, for example, about 100 electrons are freed along each centimeter of the path of a charged particle, and for each free electron a corresponding positive ion is formed. If the small amount of energy contained in this "ionization trail" can be made to produce some visible effect, the physicist can find out where the particle went. He can also measure the momentum of a particle by observing the radius of curvature of its track in a magnetic field, and he can obtain information about the way it interacts with other particles by observing sudden changes in direction of its track.

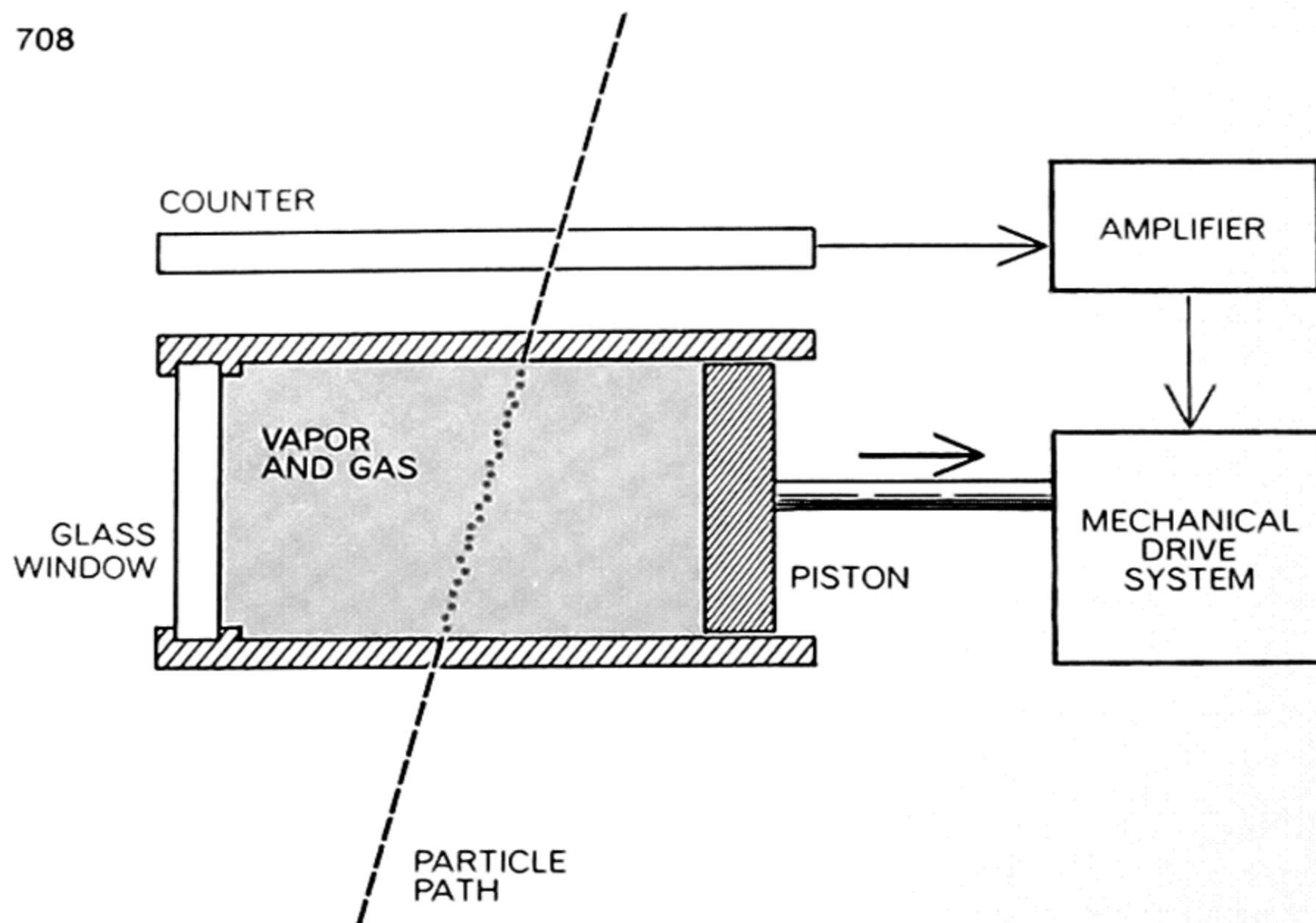
In one of the first of all elementary-particle experiments Hans Geiger and Ernest Marsden, working in the Cavendish Laboratory at the University of Cambridge, detected the small energy of an ionization trail without amplification by using the extreme sensitivity of the dark-adapted human eye. They observed the small flashes of light made when alpha particles went through certain crystalline materials called scintillators. From Geiger and Marsden's observations of the angles at which alpha particles scattered from a target into the scintillator, Ernest Rutherford concluded by 1913 that the positive charge of the atom was concentrated in a nucleus.

A fast, singly charged particle—a cosmic ray meson, for example—produces only about a thousandth as many free electrons per millimeter of track as a slow, doubly charged alpha particle

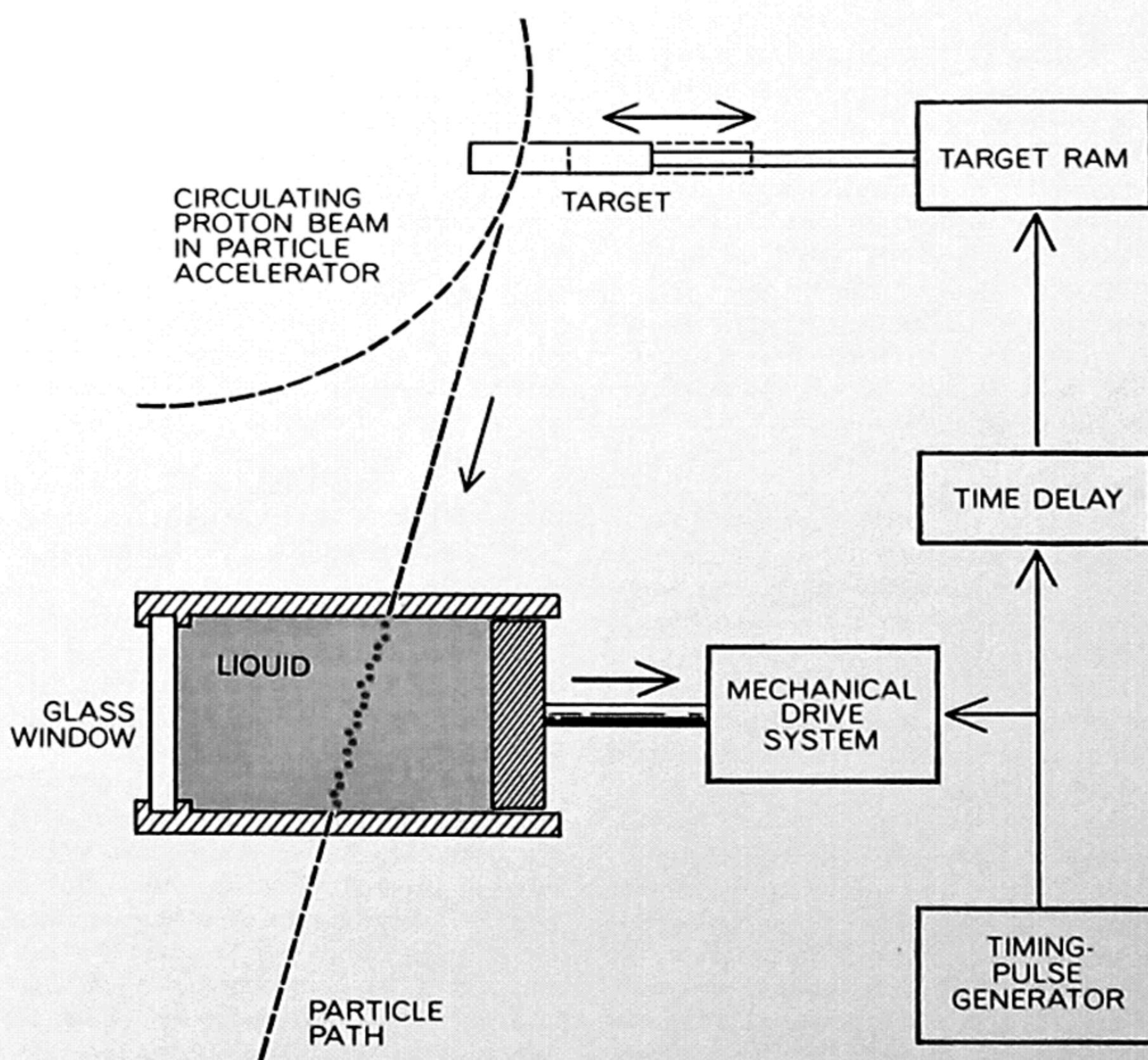
does. The detection of fast particles therefore requires some kind of amplification of the energy of the ionization trail. Since Rutherford's time the devices used to detect elementary particles have divided into two broad classes, both of which amplify. One class consists of "counters." Every counter includes a sensitive volume of gas, liquid or solid with well-defined dimensions in space. When a charged particle passes through the sensitive volume, the counter produces a brief electric pulse, or signal. The pulses can be tallied electronically; hence the name "counter."

The other class does not have a well-recognized generic name, but it can be called the class of "track detectors." A track detector shows where a charged particle went by indicating many points in space along the particle's ionization trail. Usually the information provided by a track detector is recorded by photography. In fact, for certain purposes stacks of photographic film or a single block of photographic emulsion can be used directly as a track detector. A charged particle sensitizes emulsion grains along its track and amplification is achieved by means of a chemical developer. In the next few years some advanced track detectors may be built that will put out information in the form of electrical signals.

If one compares the two classes, it is apparent that the counter gives only a limited amount of information, but it gives it immediately in a simple form suitable for direct use in electronic circuits. In modern counters the information is often available in less than 10 nanoseconds (10 billionths of a second). The track detector gives much more information, but the information goes into photographic emulsion, where it is unavailable until the emulsion is developed and analyzed. A counter with a sensitive



CLOUD CHAMBER, invented in 1911 by C. T. R. Wilson, was the first of the particle-track detectors. A counter, which simply senses the arrival of a particle, triggers the movement of a piston that expands the gas and vapor inside the chamber. This makes the vapor supersaturated, and fog droplets rapidly grow along the ionization trail left by passage of the particle. The droplets form clear tracks, which are photographed stereoscopically for analysis.



BUBBLE CHAMBER, a track detector invented by Donald A. Glaser, contains a liquid near its boiling point. When the chamber pressure is lowered, the liquid becomes superheated and bubbles of vapor grow along the ionization trail left by a charged particle. A timing mechanism moves a target into the beam of circulating protons in an accelerator, thereby directing particles into the chamber at the instant it is most sensitive to bubble growth.

volume of a cubic foot can only signal that a charged particle has passed somewhere within that cubic foot. Some track detectors with the same sensitive volume can indicate each point of the particle's path within a thousandth of a centimeter. The space resolution of the track detector balances against the reporting speed of the counter.

In modern elementary-particle experiments the experimenter often wants to trace all or part of the life histories of particles entering his detectors. He wants to identify the mass, charge and frequently the energy of each particle that enters. In addition he wants to observe if and in what way the entering particles react with the atoms in his detector. If new particles are produced by reactions, he wants to measure the properties of these product particles and to see if they decay spontaneously into combinations of other particles. In most cases, the rarer the reaction, the greater its significance. Typically only one in many thousands of particles entering a detector will produce an interesting event. If the experimenter's apparatus includes track detectors, it is much to his advantage to use counters to select those events that are worth recording in the track detector. Otherwise he may have to search through hundreds of thousands of pictures to find the rare events of interest.

The first successful track detector was the cloud chamber, invented by C. T. R. Wilson in 1911. Wilson recognized that a supersaturated vapor is unstable and that the vapor will condense into droplets around any available free ions. In cloud chambers (which are still used) a saturated vapor is maintained in a closed volume under well-controlled conditions of temperature and pressure. When a charged particle passes through the chamber, the ionization trail it leaves persists for a fraction of a second. Either before or directly after passing through the cloud chamber the particle traverses counters, which produce an electric pulse. The pulse, signaling the passage of a particle, is made to initiate the outward motion of a piston; this allows the gas inside the chamber to expand and renders the vapor in the gas supersaturated [see top illustration at left]. The vapor then begins to form droplets of fog, which condense around the ions of the charged-particle track. Droplets also tend to form around dust particles or droplets left over from a previous expansion. But under the right conditions (achieving

them is rather tricky) there forms in the chamber, in a fraction of a second, a clear trail of vapor droplets, which shows with good fidelity the path of the particle that triggered the counters. The advantage of the cloud chamber is that it can be triggered. A chamber may remain idle for hours waiting for a rare cosmic ray event, but when the event occurs and is recognized by the counters, the chamber operates on demand to record it.

Unfortunately cloud chambers have two rather serious drawbacks. First, the device is slow to set in operation, and the ionization trails persist for a large fraction of a second. As a result the number of incoming particles must be limited to prevent chamber pictures from being cluttered with more tracks than one can "read." The second drawback is the difficulty of putting into the chamber materials with which one might like to see particles interact. If material is introduced in the form of plates, the plates must be relatively few and widely spaced; otherwise the chamber will not work. If much material is needed, it must therefore be in the form of thick plates, with the result that interactions tend to occur deep in the plates, where the tracks cannot be seen. It is rather like Greek

tragedy, in which all the mayhem occurs offstage and the audience is treated only to a secondhand account of it.

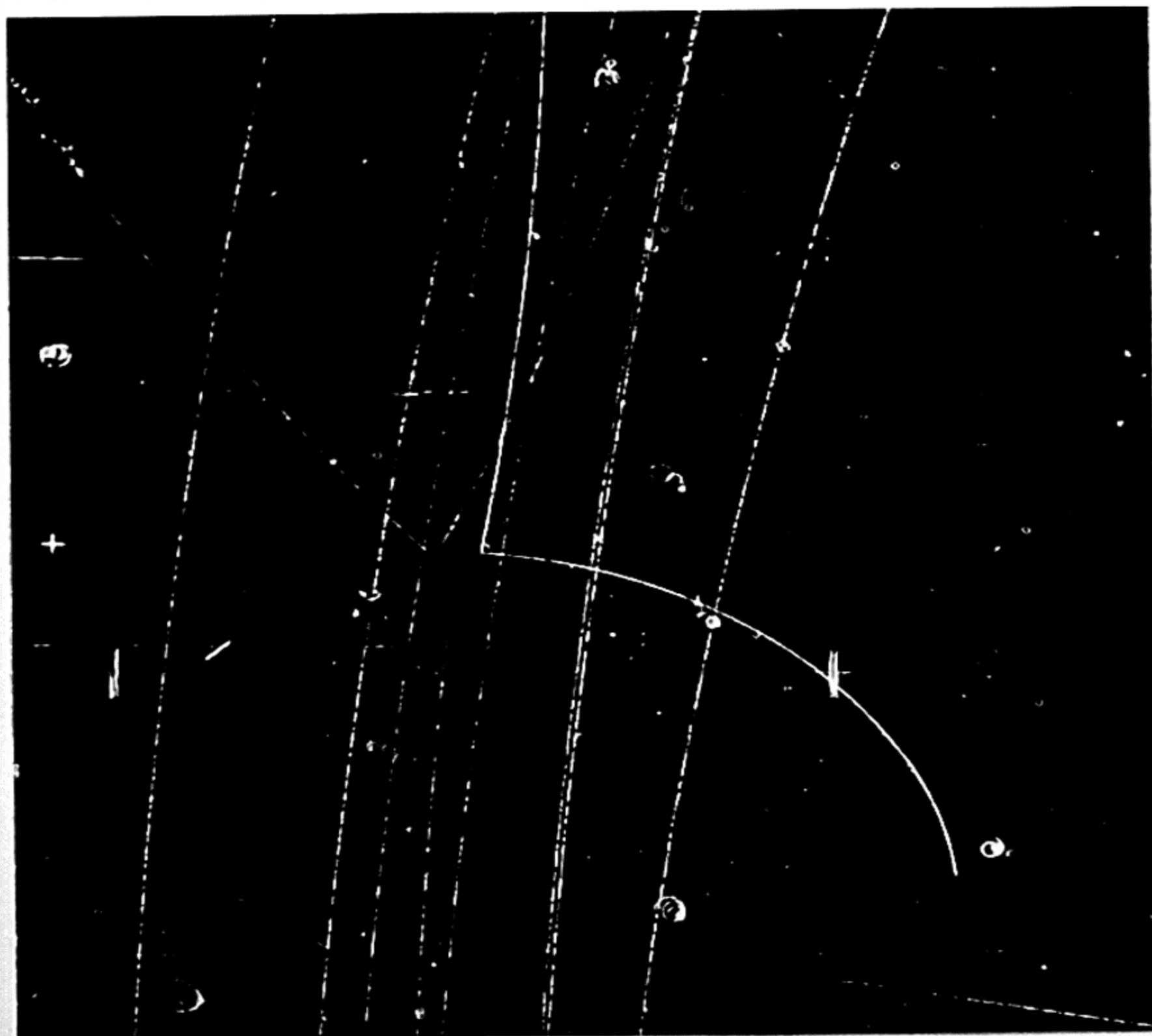
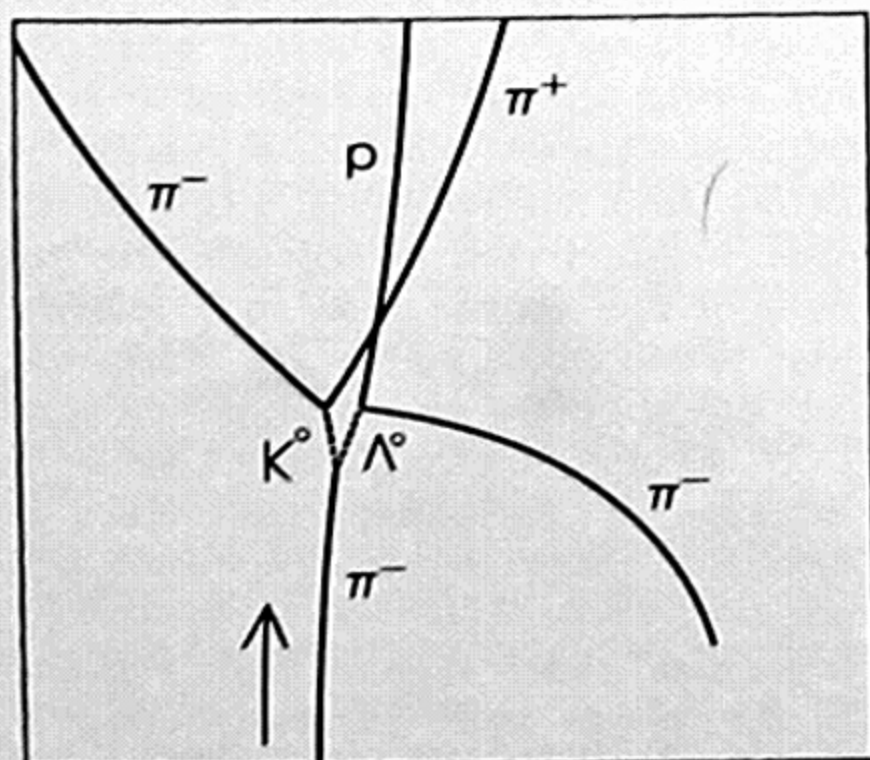
In the early 1950's Donald A. Glaser, then at the University of Michigan, developed a new type of track detector, the bubble chamber, for which he received a Nobel prize in 1960. This detector is also based on an amplification principle—the growth of bubbles in a superheated liquid. Some of the energy from an ionization trail goes into a few fast electrons, which can give up 1,000 or 2,000 volts of energy in a small volume to produce rapid local heating. If the trail is in a liquid that has suddenly been superheated by expansion, the bubbles will tend to grow fastest along the "heat track" and only slowly in other parts of the liquid. Glaser's invention was soon in use in many laboratories throughout the world, and it is safe to say that by 1959 more than half of all experimental research in elementary particle physics employed the bubble chamber.

An important virtue of Glaser's device is that one can fill the chamber with a wide variety of liquids, choosing the one that provides interactions of particular interest. For many purposes liquid hy-

drogen is ideal because it presents as a target for incoming particles only electrons and protons. In all other substances neutrons are also present. Other useful liquids are propane—in which the target atoms are carbon and hydrogen—and xenon, whose massive nucleus (54 protons and 77 neutrons) provides high stopping power. In addition the bubble chamber produces particle tracks of higher definition than those made by any other track detector, except for tracks made directly in photographic emulsion.

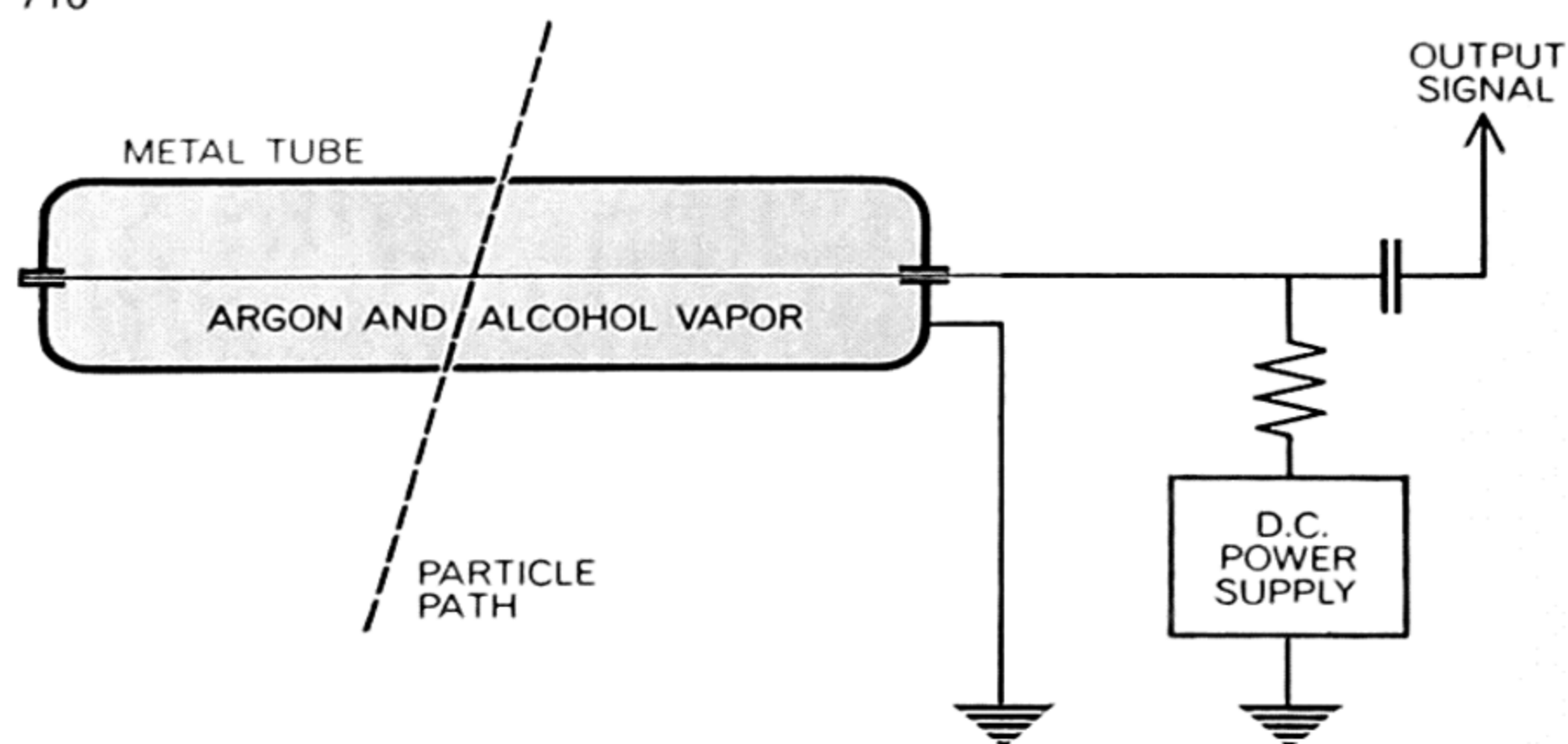
The bubble chamber shares with the emulsion method one serious disadvantage: it cannot be triggered. Since there is no way to select rare events one has no choice but to photograph the chamber at every expansion cycle, develop the films and examine hundreds or thousands of exposures looking for events of interest. Triggering is impossible because the heat track produced by a charged particle cools down in much less than a millionth of a second. This is far too short a time for the mechanical expansion system to set the chamber in operation. As a result bubble chambers are used almost exclusively with large accelerators, where a timing se-

p	PROTON
π^+	POSITIVE PI MESON
π^-	NEGATIVE PI MESON
Λ^0	NEUTRAL LAMBDA PARTICLE
K^0	NEUTRAL K MESON
e^+	POSITRON
e^-	ELECTRON
γ	GAMMA RAY

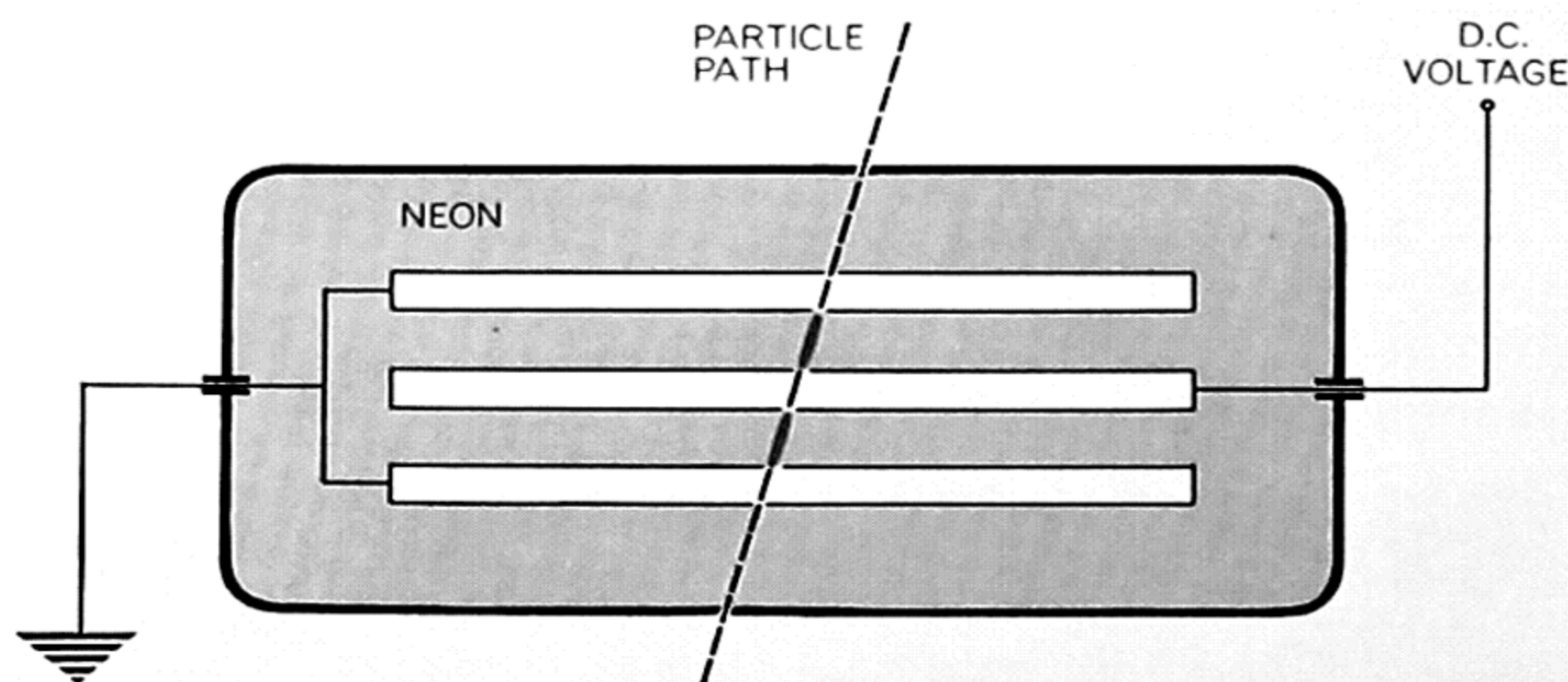


BUBBLE CHAMBER TRACKS (right) were photographed in the 72-inch liquid-hydrogen bubble chamber at the Lawrence Radia-

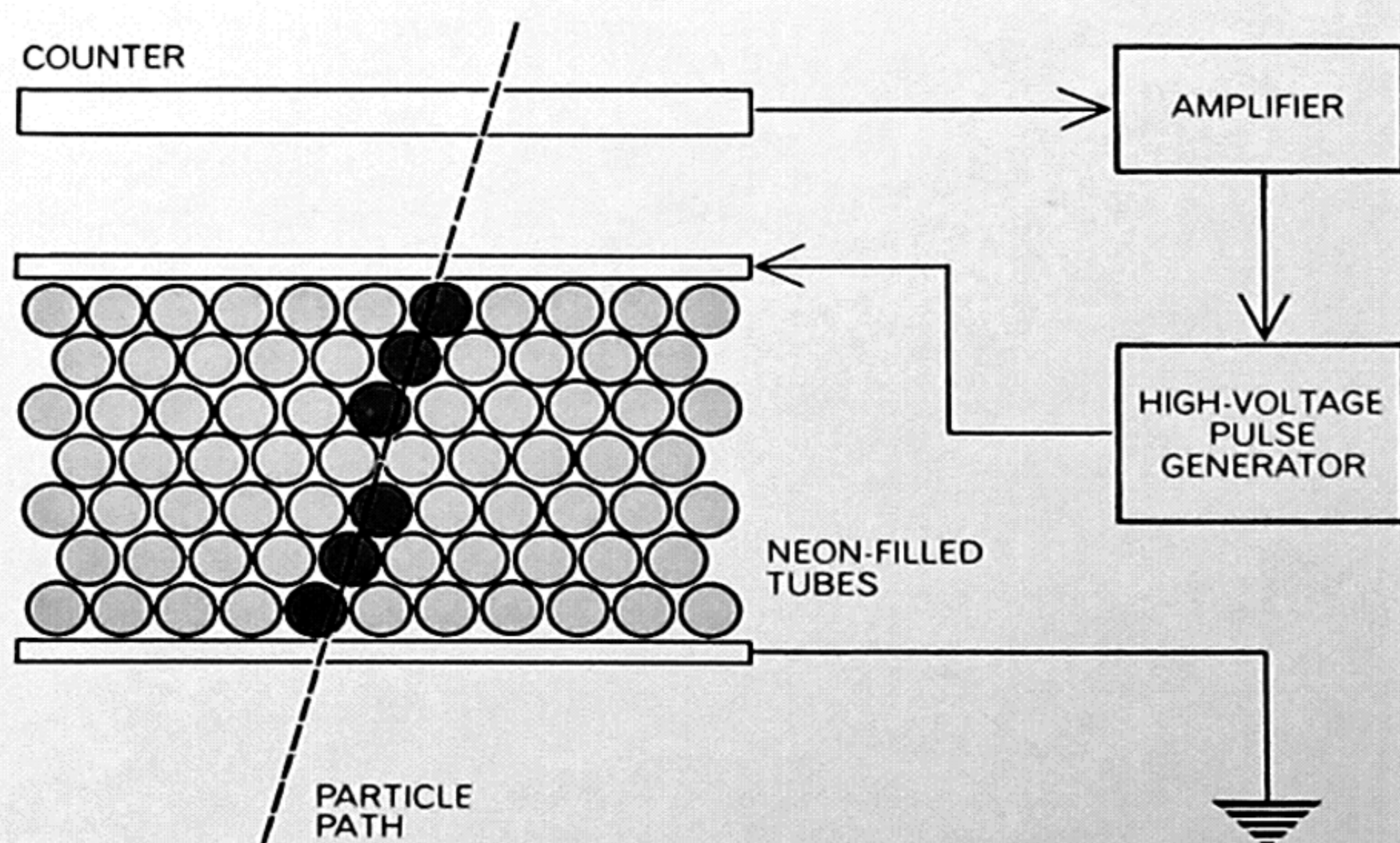
tion Laboratory of the University of California. The map and key at left identify the particles taking part in the event recorded.



GEIGER-MÜLLER COUNTER, invented in 1928, was the first device to use the amplification process available in an electric spark to detect the passage of a charged particle. A central wire inside a tube is placed at high voltage. Electrons set free from gas atoms by the passage of a particle are accelerated by the strong electric field and free other electrons in a chain reaction. The result is a large output pulse that needs no amplification to be detectable.



SPARK COUNTER was a nontriggered forerunner of the spark chamber. A high constant voltage is maintained on a metal plate placed between two grounded plates. Passage of a charged particle provides free electrons that initiate sparks in the gas between the plates.



HODOSCOPE CHAMBER, another forerunner of the spark chamber, utilizes the triggering scheme usually employed with cloud chambers. The chamber consists of neon-filled glass tubes stacked between two metal plates. When a charged particle trips the counter, a high-voltage pulse is sent to the plates, placing the tubes in a strong electric field. Tubes through which the particle passed contain ions and free electrons and therefore glow.

quence first expands the chamber, then sends in a burst of particles to be analyzed [see bottom illustration on page 708]. The chamber must then be given about a second in which to recover.

Unlike the cloud chamber and the bubble chamber, the spark chamber was the work of many hands. Its development was based on one of the most spectacular methods known for making ionization trails visible—the electric spark. The generation of an electric spark is an extremely complicated process, but it is clear that under some conditions a spark can develop from a type of chain reaction. The reaction starts when an electron from an ionized atom, accelerated by a strong electric field, bumps into and ionizes other atoms. The electrons from these atoms cause further ionizations, leading in a very brief time to a brilliant electric spark. In 1928 the amplification process available in the electric spark was used in the first of all electrical detectors for single charged particles, the Geiger-Müller counter. In this simple device, named for Hans Geiger and Walther Müller, a central wire inside a tube is charged to high voltage. When a particle goes through the counter, the electrons of its ionization track are swept toward the wire. Accelerating as they approach the wire's strong field, they ionize more atoms. The ionized atoms emit photons (light quanta), which release additional electrons from the gas, spreading the discharge. Within millionths of a second the gas all along the center wire serves as the path for an electric spark. Geiger counters make tremendous pulses, which was a great virtue when sensitive electronic amplifiers were still difficult to build.

In the 1930's the standard equipment of the elementary-particle physicist consisted of a cloud chamber triggered by Geiger counters. In the late 1940's, when Geiger counters had been generally superseded by the development of scintillation counters (faster and capable of giving more information), a few physicists began trying to use the mechanism of the electric spark in a detector that would make visible the track—not just the presence—of a charged particle. J. W. Keuffel, working at the California Institute of Technology and later at Princeton University, built several spark counters, consisting of well-polished condenser plates kept at high voltage. If the plates were carefully aligned, clean and dust-free, and maintained just below the potential needed for a spark to jump between them, they would sometimes spark preferentially along the trail

of an incoming cosmic ray particle. Keuffel suggested the use of arrays of his parallel-plate spark counters to obtain tracks of the passage of a charged particle, but these counters were so difficult to build and to operate that it was not easy to follow up the suggestion.

In 1955 M. Conversi and A. Gozzini described in the Italian physics journal *Nuovo Cimento* an intermediate type of track chamber somewhat similar to the Keuffel spark counter. Their device, called a hodoscope chamber, consisted of many neon-filled glass tubes stacked between two parallel metal plates [see bottom illustration on opposite page]. Within a few millionths of a second after the passage of a charged particle through the stack of tubes, a set of counters outside the stack triggered an electronic circuit that placed a strong electric field on the tubes. Those through which the particle had passed then glowed, much as a neon sign glows. Other tubes remained dark if the applied pulse was on for only a short time. The hodoscope chamber was fairly easy to build, and its inventors had introduced a technique that was essential for the development of spark chambers: the use of counters to pulse the electric field. In their chamber the high voltage was on only when they were sure a particle track was there to be photographed. If the high voltage had been left on continuously, as it was in the earlier spark counters, some neon tubes would eventually have fired even in the absence of an entering track. The chief defect of the hodoscope was that it revealed only two dimensions of a particle's three-dimensional path.

In 1957 two British physicists, T. E. Cranshaw and J. F. de Beer, reported in *Nuovo Cimento* the next step toward a practical spark chamber. They combined the parallel-plate geometry of the spark counter with the pulse-triggering technique of the hodoscope chamber to make an efficient spark chamber with six one-millimeter gaps. They also introduced the use of a continuous electric clearing field to remove from the chamber ionization trails older than a few microseconds. This electric field, well below the threshold needed to make a spark, caused a slow continuous drift to the plates of all free electrons and ions released in the chamber gas. In this way it "erased" ionization trails in a few microseconds. A similar clearing field had long been used in cloud chambers to sweep out the slow-moving positive ions.

It happened that Cranshaw and de Beer chose to use air rather than neon in their chamber, and this small difference made it impossible for their chamber to

detect two or more simultaneous tracks. Still, their work was so successful that several other groups—in Germany, Japan, the U.S.S.R. and the U.S.—continued to work along similar lines.

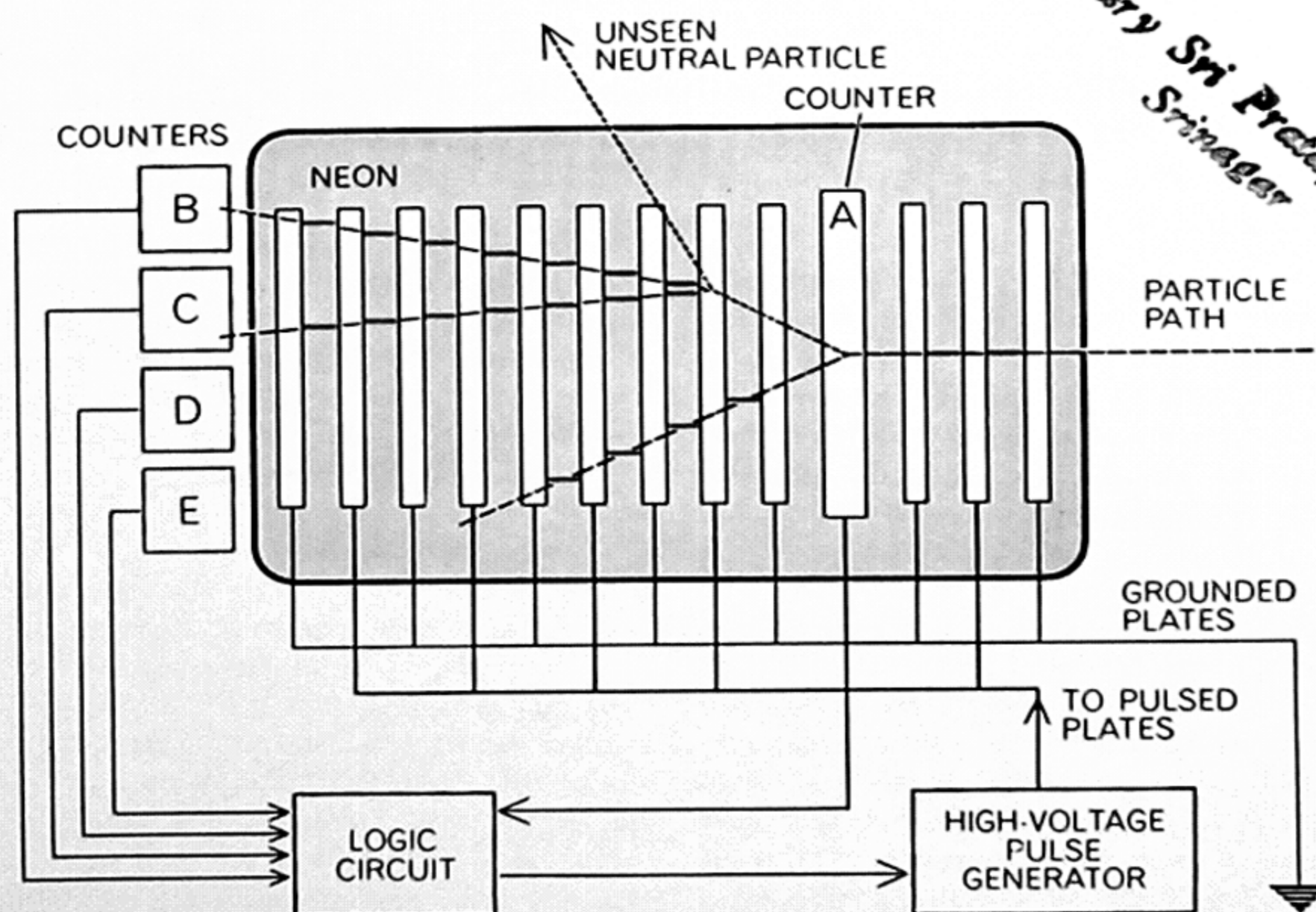
The final step—substitution of neon for air—was taken by S. Fukui and S. Miyamoto of Osaka University and reported in 1959. The two Japanese physicists were interested in developing a track detector that could be used for cosmic rays. Bubble chambers are not useful for such work, since they cannot be triggered. Fukui and Miyamoto found that in a chamber containing neon rather than air several simultaneous particle tracks could be seen.

One big difference between the behavior of air and of neon in spark chambers is that oxygen molecules (O_2) in air can combine with the free electrons of the ionization trail, whereas neon atoms cannot. The inertness of neon—and of other "noble" gases—is explained by the fact that it has a full complement of eight electrons in its outer electron shell. In contrast an oxygen molecule can acquire one electron and thereby become a negative ion (O_2^-). The electrons are well anchored to the oxygen molecules, some 60,000 times more massive than themselves, and cannot be freed except by application of a strong electric field.

Consequently an air-filled spark chamber requires an operating pulse of 7,000 to 10,000 volts for each millimeter of space between its plates. This is about three times the voltage needed for a neon spark chamber.

The formation of oxygen ions also explains other characteristics of an air spark chamber. If the electron in an ionization trail can migrate freely to the plates of the chamber, its travel time is brief. But if it is attached to an oxygen molecule along the way, the velocity of the resulting ion is much slower than that of the electron. In fact, if the mass of a particle is suddenly increased by 60,000 times, its velocity must decrease by the square root of 60,000, or by a factor of about 250. Because most of the electrons liberated in an air spark chamber are slowed down in this fashion, they require many microseconds to migrate to the plates of the chamber. Such a chamber therefore remains sensitive for a long time, and in it old tracks cannot be quickly erased.

It is not so clear why air chambers show only one spark per gap even though several ionization trails may be present. It may be that at the high electric fields needed to operate such chambers the spark produced by the first electron freed from an oxygen ion occurs so rapidly that the plates are quickly dis-



SPARK CHAMBER, which became practical with the work of S. Fukui and S. Miyamoto in 1959, consists of an array of thin metal plates surrounded by neon. It is also provided with counters and a "logic" circuit for determining when a particle meeting certain criteria has appeared. When it appears, a high-voltage pulse is sent to alternate plates and sparks occur along the ionization trails left by each charged particle. In the example shown, a charged particle interacts in counter A, yielding one neutral and one charged secondary. The secondary decays in the chamber, producing two charged particles and a neutral one.

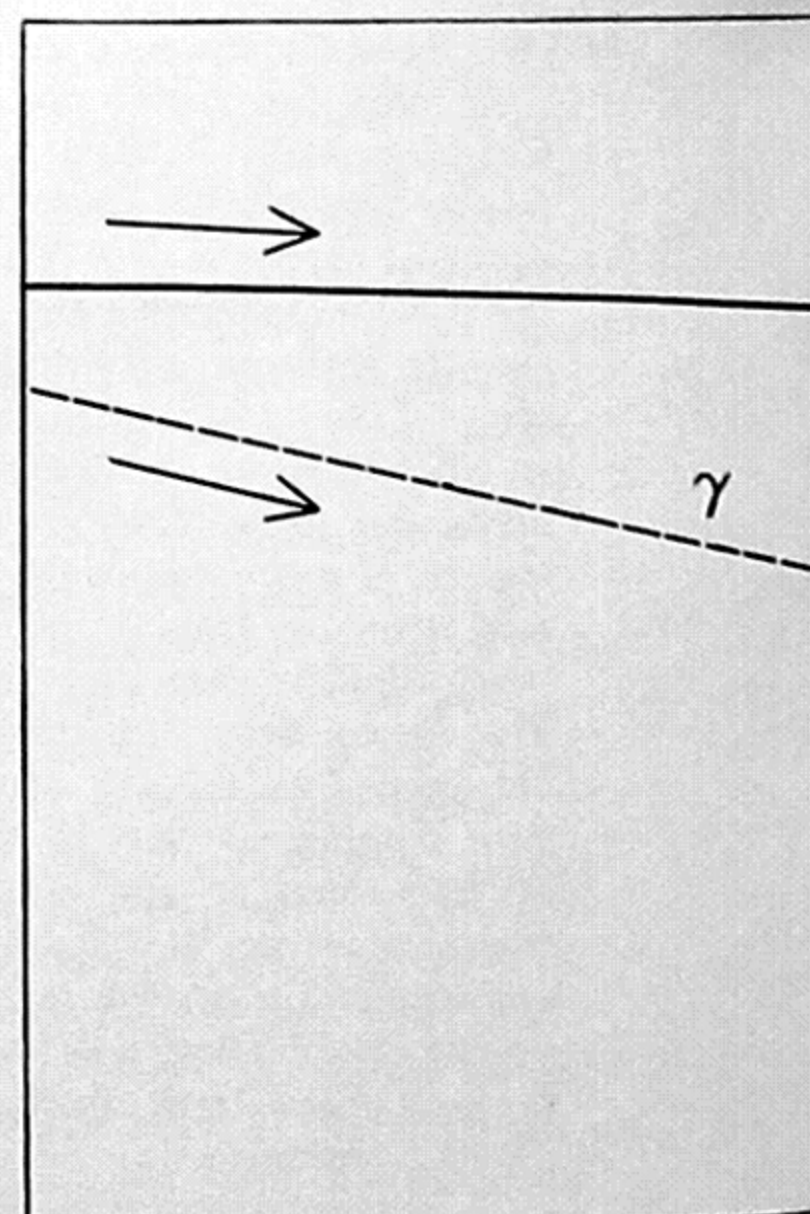
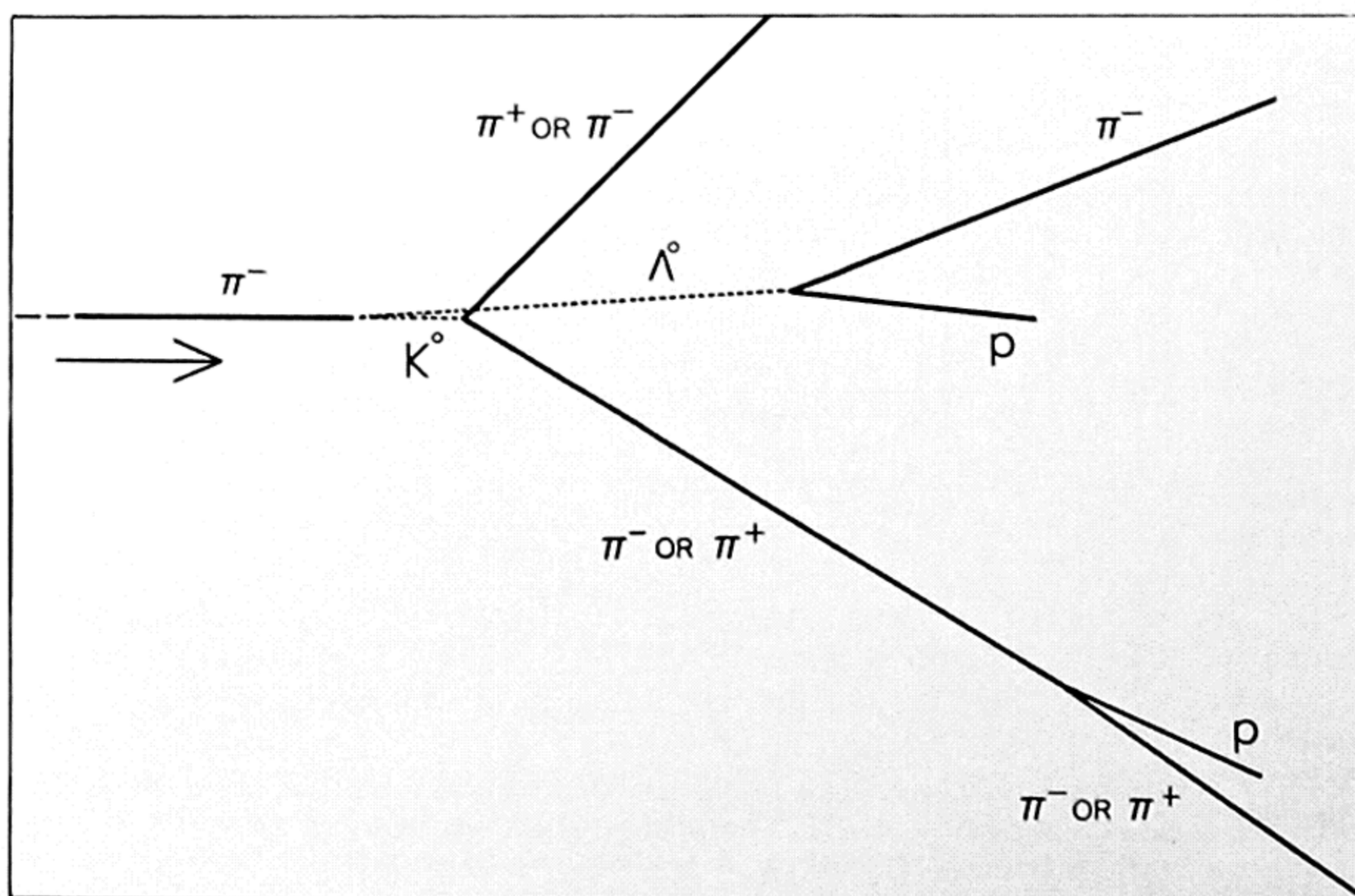
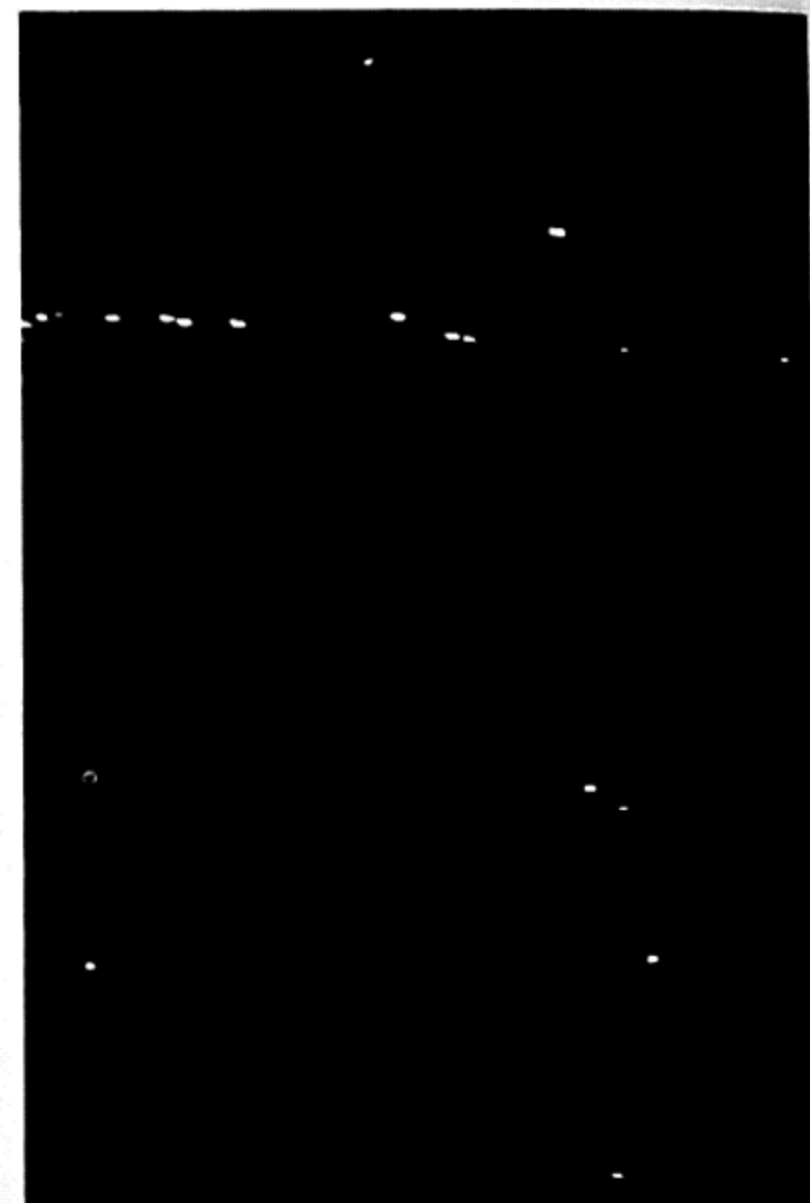
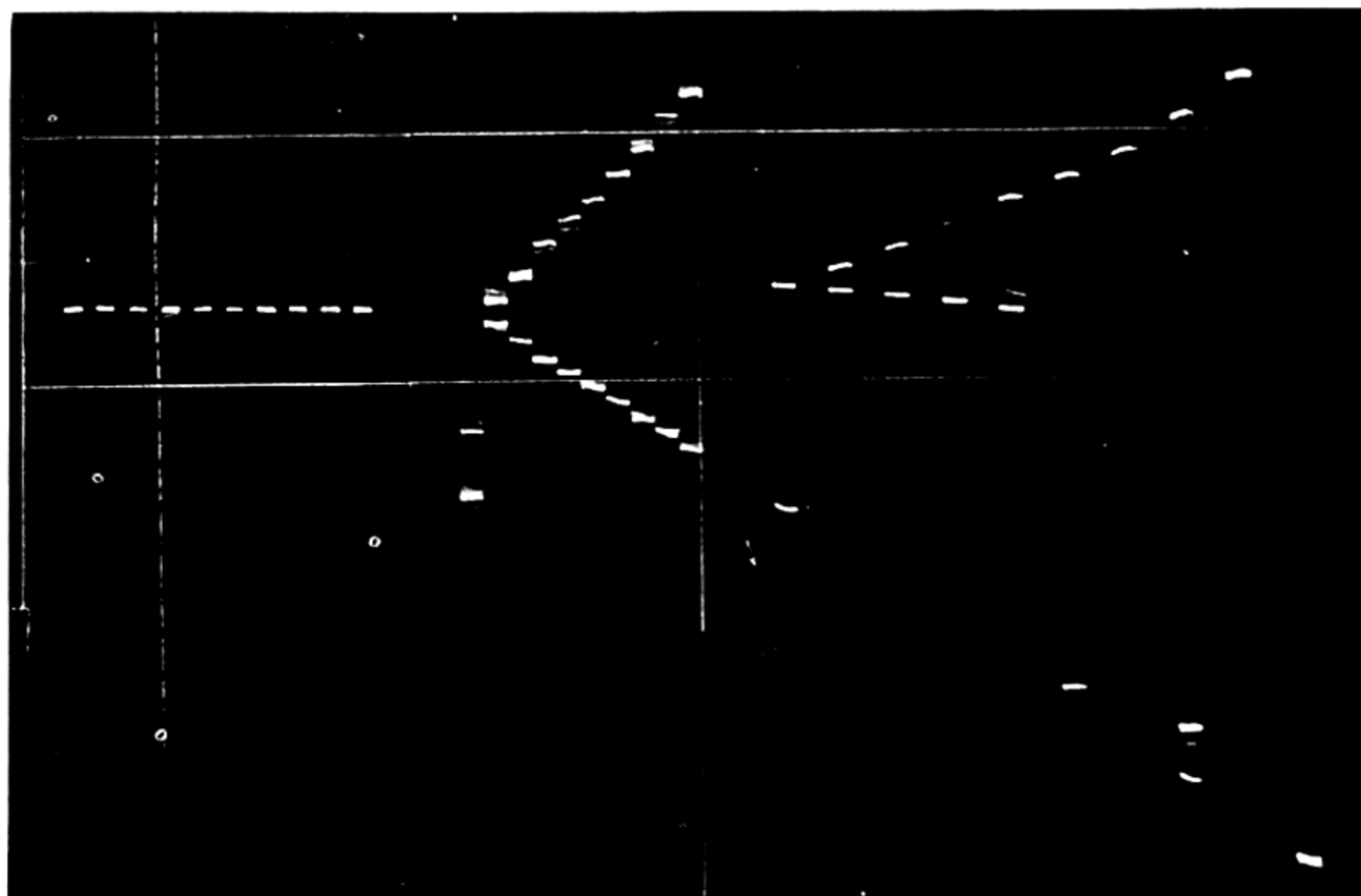
charged below the threshold field, preventing any other attached electrons from getting free to start other sparks. This is consistent with an observation by Cranshaw and de Beer that only one electron is needed to start the spark.

Following the announcement of a practical spark chamber by Fukui and Miyamoto in 1959, the idea was immediately taken up by physicists in the U.S. and elsewhere. Within a matter of months Bruce Cork of the University of California had built a six-gap spark

chamber and had operated it in a beam of particles from the six-billion-electron-volt accelerator of the Lawrence Radiation Laboratory. Almost simultaneously James L. Cronin of Princeton University built and operated a large 18-gap spark chamber, which yielded high-quality pictures of the tracks made by cosmic rays and by accelerator-produced particles. Both of these chambers used noble gases (neon or argon) and employed clearing fields to erase the ionization trails. Cork and Cronin were also the first to conduct actual experiments using

a spark chamber as a particle detector.

In their work, as in most subsequent experiments using spark chambers, the occurrence of an interesting event was recognized by a system of conventional counters, which then triggered the operation of the chamber. Typically particles arrived at the spark chamber at intervals of a few microseconds and their tracks were swept to the plates by the continuous clearing field after only one microsecond. Consequently the pulsing of the spark chamber had to be carried out in much less than one microsecond so that



SPARK CHAMBER PICTURES show the appearance of particle tracks when the particles are curved by a magnetic field (*top right*) and when they are not (*top left*). The maps below each picture

identify the charged particles, which leave tracks, and the neutral ones, whose presence is inferred. The reaction at the left was seen in a spark chamber operated at Brookhaven by James L. Cronin

the interesting track would still be there to be detected by spark amplification.

Within the past three years a wide variety of spark chambers have been built, each designed to exploit certain desirable features. Some have been made with thick carbon plates to allow interactions of the incoming particles with carbon. Others have been built in the form of a cylinder, to study the scattering of particles by a target located on the axis of the cylinder.

Along with several other physicists, I have been particularly interested in the

design and use of thin-plate spark chambers that can be operated in a magnetic field. In a uniform magnetic field the path of a charged particle of constant energy is a circle whose radius is proportional to the momentum of the particle. The idea of using a magnetic field to obtain momentum information goes back to the early days of the cloud chamber, and bubble chambers are nearly always operated in such a field. The measurement of the momentum of each charged particle in a reaction is always useful, and frequently essential, for identifying the particles and learning the details of their interactions.

When a magnetic field is used in a spark chamber, the sparks trace the ionization trails more closely if the spacing between the chamber plates is small. As the spacing is reduced, however, it becomes increasingly important for the plates to be flat and uniformly spaced, and the triggering pulse has to rise from zero to the peak voltage at higher speed. Fukui and Miyamoto had used spacings of 10 millimeters. Cork's chamber had a six-millimeter spacing. Within a few months we found in our laboratory at Princeton University that the spacing between spark-chamber plates operated in neon could be as small as two millimeters.

Unless very close plate-spacing is wanted, the construction of a spark chamber is not too difficult and might make a feasible project for an amateur scientist. A chamber with an adjustable plate spacing of two to 10 millimeters, the first model built by our group, was largely the work of college sophomores majoring in physics. Our second instrument was small but operated in a magnetic field. It contained 50 gaps of three millimeters each, separated by aluminum foil a thousandth of an inch thick. A third chamber, with 128 gaps of three-millimeter spacing and a volume of two cubic feet, can measure the momentum of particles with good accuracy. When the tracks cross 100 or more gaps, the accuracy of momentum measurement approaches that obtainable in a good bubble chamber.

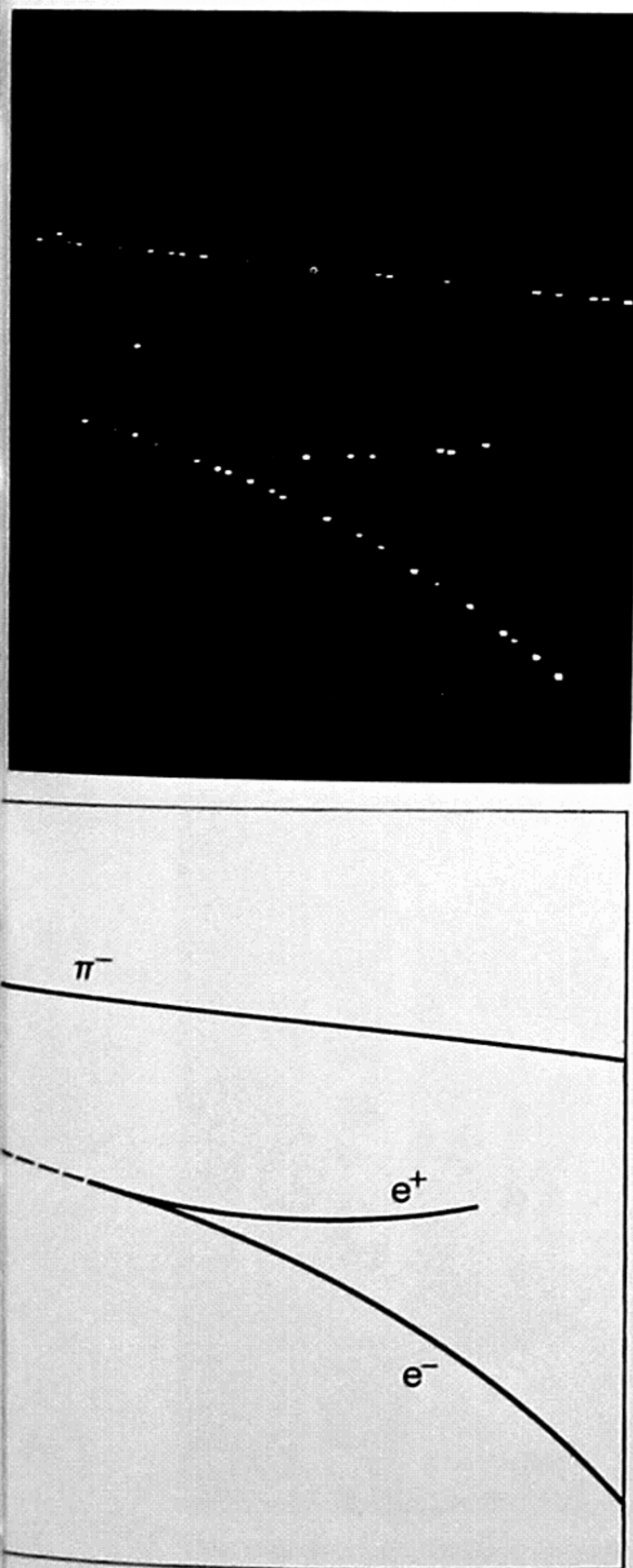
At present the advantages the bubble chamber retains over the spark chamber are two. First, pure liquid hydrogen can be used as the only material in the bubble chamber, thereby limiting nuclear reactions to those between elementary particles and hydrogen nuclei (protons). In 1960 we studied the possibility of imitating a hydrogen bubble chamber by using liquid-hydrogen-filled hollow plates in an atmosphere of gaseous helium. We established that

such a chamber would work but so far no one has needed its properties badly enough to build one. The second advantage of the bubble chamber is that it yields very fine ionization trails, and it produces them no matter which way the particle is moving. The bubbles trace a particle's path with an uncertainty of less than a thousandth of an inch. Even in narrow-gap spark chambers the sparks scatter in a region 15 to 20 thousandths of an inch wide. Moreover, in a spark chamber the path uncertainty increases as the particle approaches a course parallel to the plates.

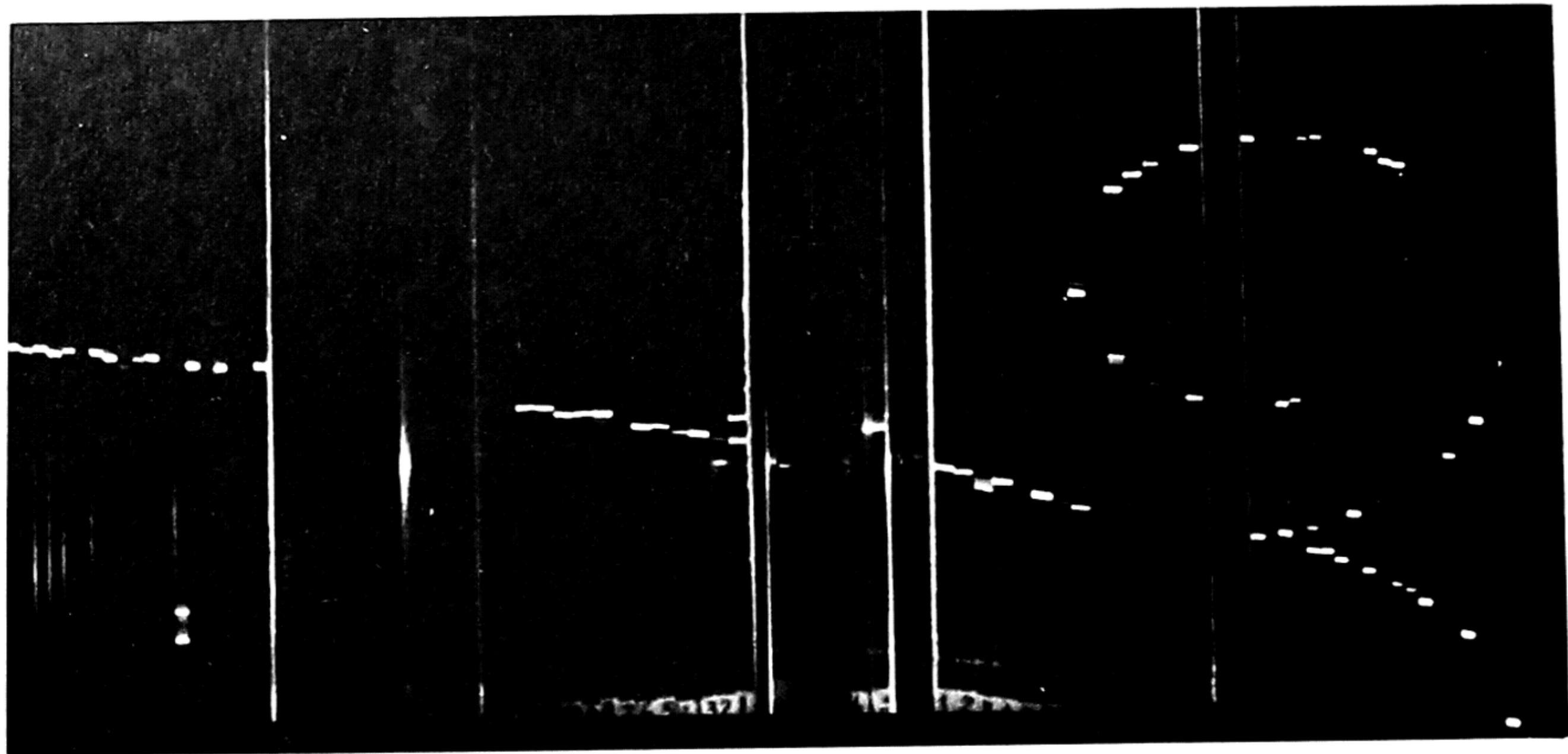
In spite of these drawbacks the spark chamber has two big advantages over the bubble chamber. First, the decision to photograph a given event can be made after the event has occurred. Second, because old ionization trails are swept to the walls after only one or two microseconds the spark chamber picture shows only the tracks produced during the last microsecond before the chamber was pulsed. Because of these two features one can select and photograph an interesting event caused by a single entering particle out of many thousands, all arriving over a few thousandths of a second. Each ionization trail of the uninteresting majority of tracks is swept away and does not remain to confuse the picture.

The decision as to which events to photograph is made by "logic" circuits that analyze the output of counters, which may be located outside or inside the spark chamber itself. Frequently the logic requirements are severe and the pulses from many counters must be digested and analyzed before a decision is made whether to pulse the chamber or not. Ordinarily a time of about 100 nanoseconds (100 billionths of a second) is available for the decision. This is not uncomfortably short with present-day circuitry. For the past 10 years it has been practical to use circuits that operate in 20 nanoseconds or less.

Those of us who have jumped on the spark chamber bandwagon are naturally enthusiastic about future prospects for the instrument. We have found that physicists who formerly used bubble chambers are delighted to have a device that eliminates great masses of uninteresting pictures. And former counter physicists are happy to see the tracks they knew were going through their counters. We all know that neither bubble chambers nor counters are going to be put out of business by the new track detectors, but to a remarkable degree spark chambers allow us some of the best of both worlds.

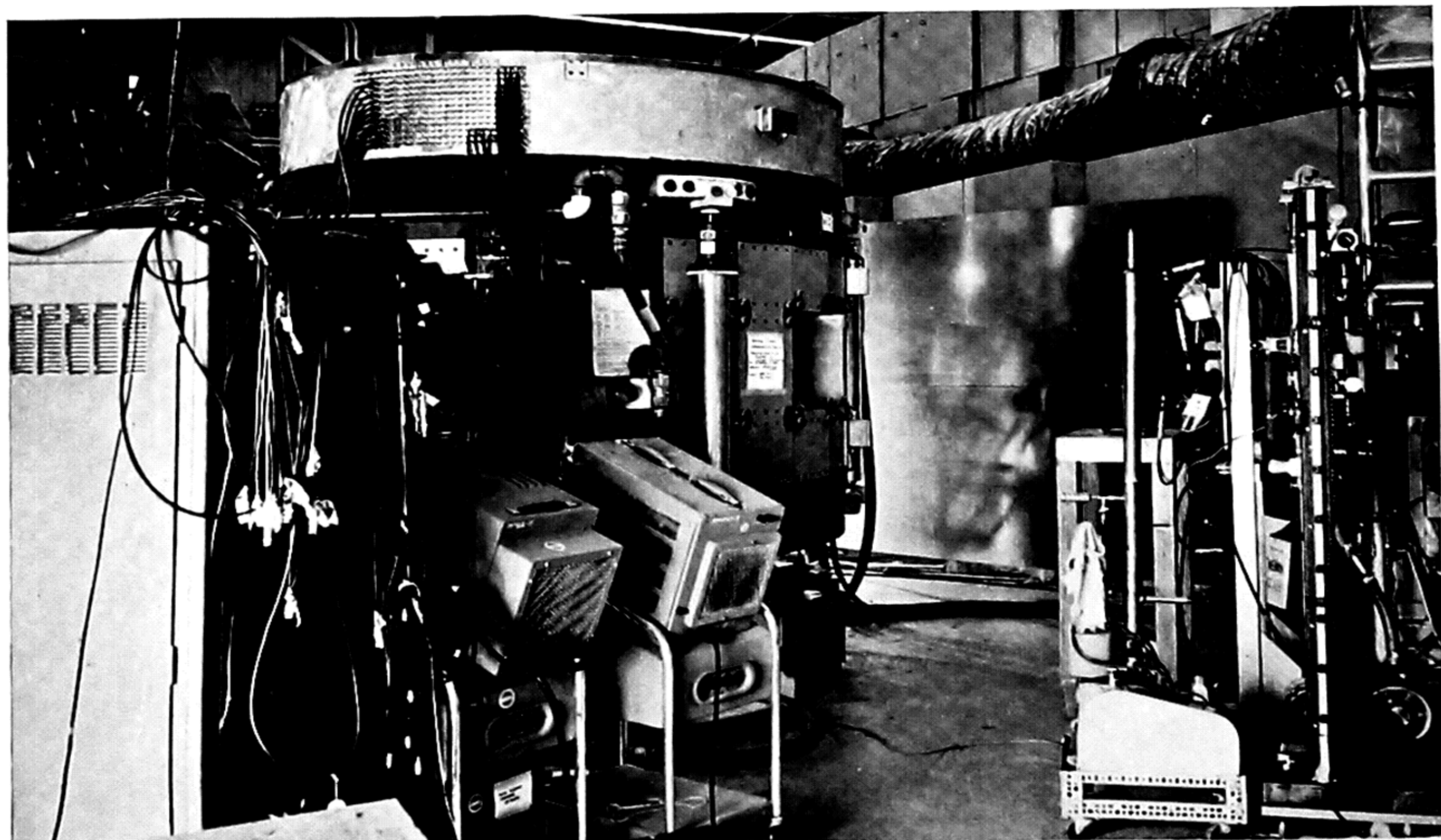


of Princeton. The picture at right was made in author's two-cubic-foot spark chamber at Brookhaven, shown at bottom of page 714.



TRACKS OF CHARGED PARTICLES are made visible by spark chamber operated at the European Organization for Nuclear Research (CERN) in Geneva. The chamber contains a series of metal plates surrounded by neon, in which a charged particle leaves an ionization trail. If a high-voltage pulse is applied quickly to alternate plates, sparks follow the trail. Here a negative pi meson enters

at left and reacts to give an invisible neutral pion, which decays into two invisible gamma rays. One gamma ray yields the tracks of an electron and positron, which are curved in opposite directions by a magnetic field. The chamber was built by Arthur Roberts, G. R. Burleson, T. F. Moang, P. Kalmus, L. Niemala and T. A. Romanowski of the Argonne National Laboratory and B. Leontic of CERN.



SPARK CHAMBER INSIDE MAGNET at Brookhaven National Laboratory was built by the author, F. V. Murphy, Richard B. Wattson and Kenneth E. Wright of Princeton University. The diameter of the magnet is eight feet. The spark chamber has a volume of two cubic feet and contains 128 metal plates spaced three millimeters

apart. The magnet makes charged particles curve as they pass through the chamber. The particles originate in the 30-billion-electron-volt proton accelerator, beyond the wall at right, and enter through the port at the right side of the instrument. A photograph made by this device appears at upper right on pages 712 and 713.

The Author

GERARD K. O'NEILL is associate professor of physics at Princeton University. After spending two years in the Navy (1944–1946) he entered Swarthmore College and received a B.A. in 1950. After obtaining a Ph.D. in physics in 1954 from Cornell University he began teaching at Princeton. He spent most of the next five years helping in the design and construction of the three-billion-electron-volt proton synchrotron being built jointly by Princeton and the University of Pennsylvania. O'Neill became interested in building spark chambers in 1960. He is now preparing spark chamber experiments for the three-Bev synchrotron, which is due to start operation this year.

Bibliography

- ELECTRICAL BREAKDOWN OF GASES. J. M. Meek and J. D. Craggs. Oxford University Press, 1953. See pages 251–290.
- DAS ELEKTRONENZÄHLROHR. Hans Geiger and Walther Müller in *Physikalische Zeitschrift*, Vol. 29, No. 22, pages 839–841; November 15, 1928.
- A NEW TYPE OF PARTICLE DETECTOR: THE "DISCHARGE CHAMBER." S. Fukui and S. Miyamoto in *Nuovo Cimento*, Vol. 11, No. 1, pages 113–115; January 1, 1959.
- A TRIGGERED SPARK COUNTER. T. E. Cranshaw and J. F. de Beer in *Nuovo Cimento*, Vol. 5, No. 5, pages 1107–1117; May 1, 1957.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.



NEUTRINO DETECTOR ASSEMBLY was designed by Frederick Reines of the Case Institute of Technology to measure the flux of high-energy neutrinos in cosmic rays. It is photographed in position in the Fairport Harbor Mine of the Morton Salt Company near

Cleveland. The detector is located 2,000 feet underground to shield out particles other than neutrinos. The tank in foreground, which is the detector proper, is eight feet eight inches high and eight feet in diameter. Structure behind it houses electronic apparatus.

NEUTRINO ASTRONOMY

by Philip Morrison

A flood of the penetrating particles called neutrinos pours down on the earth from the sun and stars. If they could be detected, they would constitute a rich source of astrophysical information.

Of the dozens of "fundamental particles" with which physics is seized, only two nowadays bear the name without any skeptics' quotation marks. These are the photon and the neutrino. Each stands at the base of one of the two hierarchies to which all known particles belong. The photon, or quantum of the electromagnetic field, is the simplest and lightest of the bosons: particles with an integral number of spin units. The neutrino occupies a corresponding place among the fermions: particles with half-integral spin. (The familiar proton, neutron and electron are all fermions.) Neither the photon nor the neutrino possesses a measurable mass, and they have energy only by virtue of their motion, which is always at the speed of light. The photon is familiar to everyone; all the light by which we live and know the world is nothing but a flood of photons. The neutrino lies almost entirely outside human experience. Yet it would be strange if nature's other most fundamental particle did not conceal a fair share of nature's secrets. Recent theoretical calculations indicate, in fact, that the unfamiliar neutrino may turn out to be the major actor in an unseen drama of the physical world.

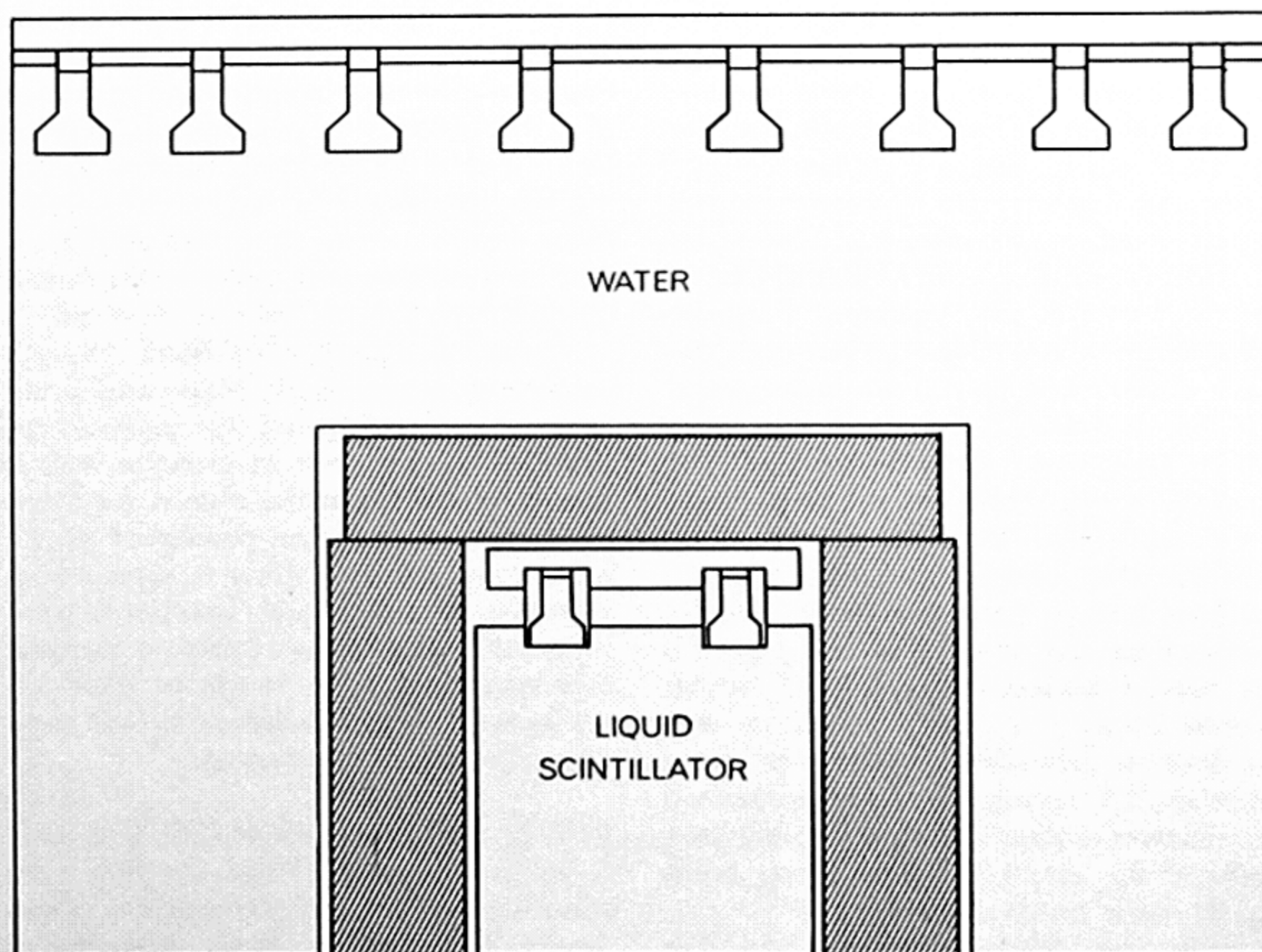
Everything the astronomer knows about the regions beyond the earth he has learned from photons. The faintest star that can be seen with the naked eye rains some 1,000 photons per second into the wide-open, dark-adapted pupil. Sunlight delivers to the same area a million billion times as many. The signals picked up by the radio astronomers' huge antennas are also photons, but photons of much lower energy.

Neutrinos, it is now realized, must also stream down on us out of the heavens. From the sun they bring a tenth as much energy as does the visible light.

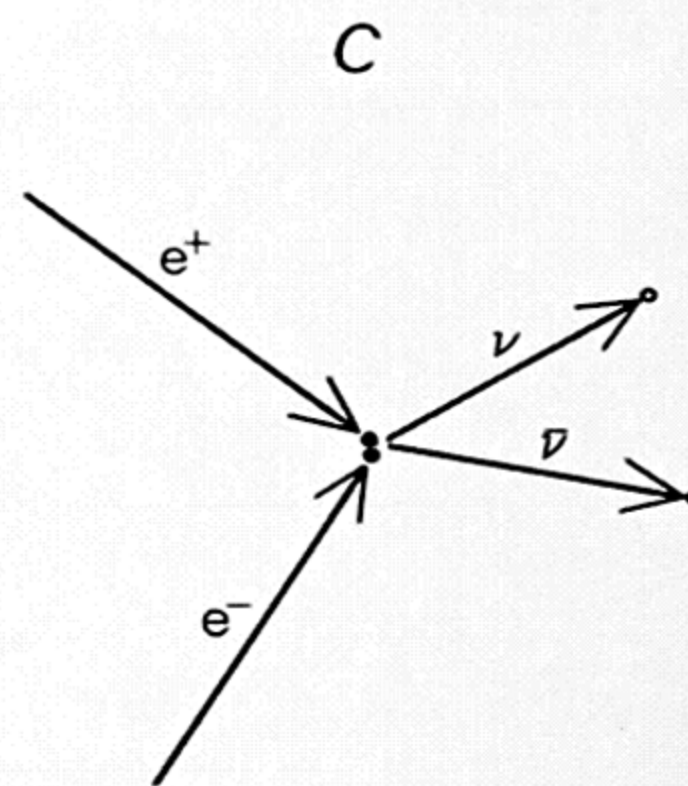
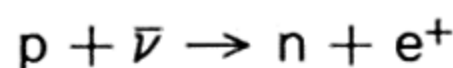
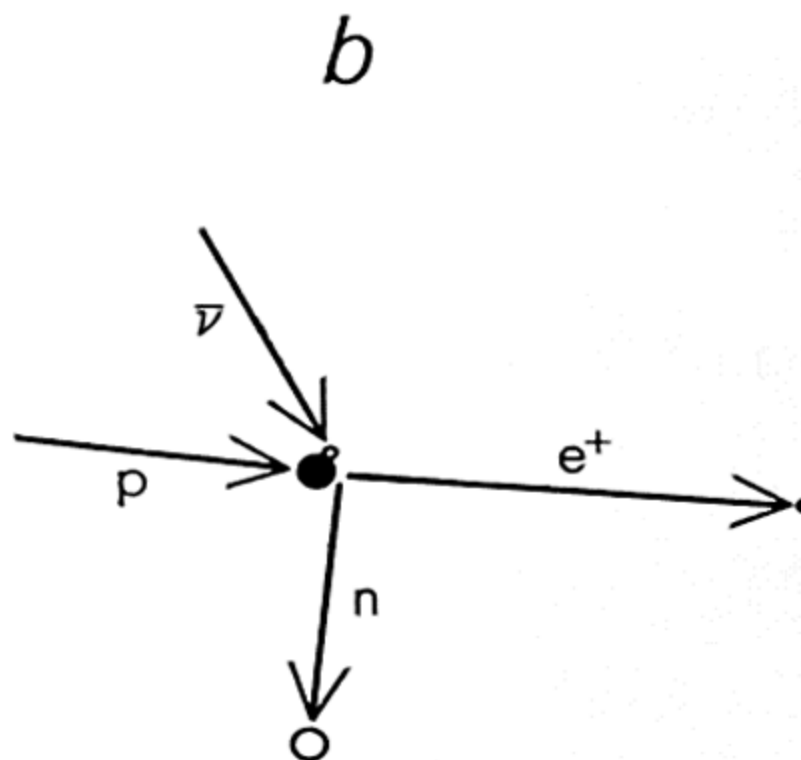
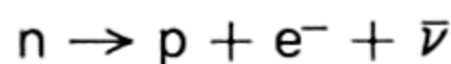
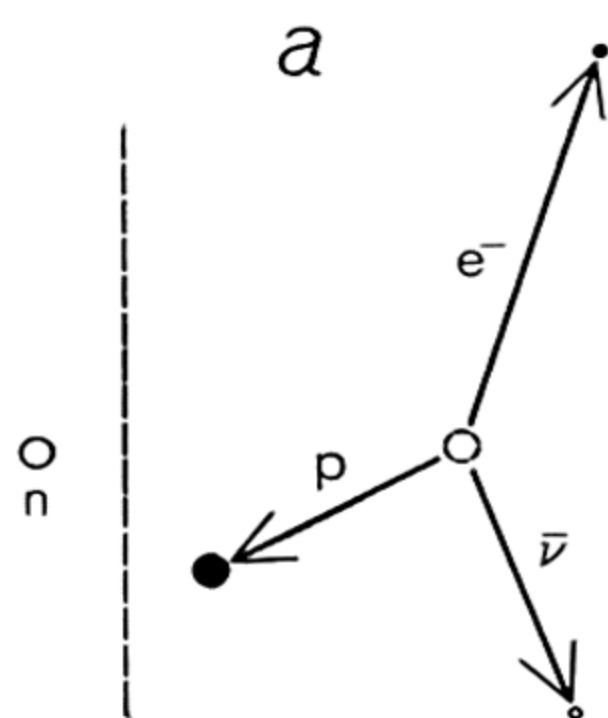
From many stars they convey vastly more energy than comes in the form of light. Yet this enormous flux of particles passes wholly unnoticed. Suppose that somehow we could build a neutrino telescope. What could we expect to see of the sun and stars? What might we hope to learn of the universe? This article will sketch some of the hopes.

The task of a would-be builder of a neutrino telescope is not to gather or concentrate the particles; there are plenty of them. What he must do is in-

duce them to interact with the material of his instrument and deliver up their energy, as photons deliver theirs to the retina of the eye or to the photographic emulsion. The catch is that neutrinos interact practically not at all—with any kind of matter. At the most conservative estimate 10^{23} neutrinos pass through the body of a man during his lifetime. *Just one* is likely to interact and be absorbed there. Neutrinos from the sun come up through the solid earth at midnight as plentifully as they rain down from the



CROSS SECTION OF DETECTOR is shown schematically. The liquid scintillator responds to products of reactions between incoming neutrinos and nuclei in surrounding iron plate (*hatching*) or in scintillator itself. Scintillations are detected by photomultiplier tubes at top of inner chamber. If scintillation is accompanied by visible Čerenkov radiation in the surrounding water, the event is not counted because it was produced by a charged particle.



NEUTRINO REACTIONS are characterized by their extremely slow rate. Of the three reactions shown here, *a* and *b* have been observed, whereas *c* is known only from theory. In *a* a free neutron (*n*) disintegrates spontaneously into a proton (*p*), an electron

(*e*⁻) and an antineutrino ($\bar{\nu}$). In *b* a proton and an antineutrino ($\bar{\nu}$) combine to form a neutron and a positron (*e*⁺). In *c* an electron and a positron combine to produce a neutrino and an antineutrino. This reaction is believed to occur in stars that become supernovae.

noonday sun. How then do we know them at all?

A generation ago the neutrino made its debut, not in the laboratory but in the mind of the theoretical physicist Wolfgang Pauli. He was concerned with the problem of beta decay: the emission of electrons by certain radioactive species of atomic nuclei. Quantum theory indicated that every nucleus of a given species has precisely the same set of discrete energy states. A beta ray electron carries away some nuclear energy, and accordingly all electrons ejected from nuclei in a given state were expected to have the same amount of energy. This was indeed true of the other two kinds of radioactive emission: alpha particles and gamma ray photons. But it was not true of the electrons. They exhibited not a single energy for each decaying nuclear type but a whole spread of energies, from zero up to the maximum that was characteristic of the nuclear state. There were three ways to account for this: (1) the product nucleus for each individual electron emitted had an energy content that depended on the energy of the electron its parent nucleus had emitted, which would imply that one electron or one nucleus is intrinsically different from another, (2) energy was not conserved in electron emission, (3) a varying portion of the energy of the decay went off in some undetected form.

Only the third alternative was less than revolutionary. So physicists, always conservative, began to look for the missing energy. It had to be carried in a neutral particle, since there was no missing charge. A gamma ray was the first guess,

but it was simply not there. It was Pauli who realized that the comparative rates of electron and gamma emission furnished a clue to the elusiveness of the hypothetical particle. The release of energy by beta decay is unimaginably slower, or less probable, than the release of energy by gamma rays. The ratio is something like 10¹⁸ to one. If a new, light neutral particle were created with each beta ray electron, this particle would be almost impossible to detect: the slowness of its emission implied an analogous and matching improbability of absorption in the nuclei of any matter it penetrated. Enrico Fermi made detailed calculations showing that the shape of the electron spectrum agreed with these assumptions, and he gave the new particle the name neutrino (in Italian "little neutral one").

Physicists sighed with relief: energy conservation was safe. More and more indirect evidence piled up: momentum too was missing, but in just the way it would be missing if the elusive neutrino bore it away. Still, the profession was a bit uneasy. It is one thing to satisfy conservation laws; it is quite another to have a particle with some reality, a particle that could again be made to yield up the energy and momentum it had been given license to steal away.

It was not until modern skill and scale became available that the first neutrino was caught and the validity of the theory founded by Pauli and Fermi proved in a direct way. Six years ago Frederick Reines and Clyde L. Cowan, Jr., then at the Los Alamos Scientific Laboratory, set up an enormous detector and counter apparatus in the avalanche

of neutrinos coming from one of the world's largest nuclear reactors in Savannah River, Ga. The neutrino density there was perhaps 30 times larger than the density of neutrinos from the sun and stars.

Once or twice an hour, on the average, the experimenters caught a neutrino interacting with a proton in the huge tanks of water of their detector. Then and there appeared the lost energy and momentum of beta decay, as had been calculated.

In June a second type of neutrino was detected. It is born not with an electron in beta decay but with a mu meson in the decay of a pi meson. I mention it here only to rule it out of the discussion. This article is concerned exclusively with the beta-decay neutrino.

In addition to its original function as a carrier of energy and momentum, the neutrino has come to play a wider role in the schemes of the theoretical physicist. There seems little doubt that all fermions (and their corresponding antiparticles) can emit or absorb a neutrino (or antineutrino) in any reaction that obeys the various conservation laws of physics: conservation of energy, of momentum, of electric charge and a few more. In the process any pair of fermions can turn into any other pair. Such transformations have a small but definite rate, depending very much on the energy released but not at all on which particular fermions are involved. Another rule of this particle game is that the birth, or emission, of any antiparticle is exactly equivalent to the death, or absorption, of the corresponding particle and vice versa.

As an example of a fermion trans-

formation, consider the case in which a neutron and a neutrino are converted into a proton and an electron. In the shorthand of the particle physicist, in which the neutrino is designated by the Greek letter ν (nu), we can write

$$n + \nu \rightarrow p + e^-$$

The equation states that a neutron absorbs a neutrino and changes to a proton and an electron. (The electron carries a minus sign to distinguish it from its antiparticle, the positron, which is written e^+ .) Although the conversion is one of the most common of all neutrino reactions, it is usually not seen in this form but in the exactly equivalent form where the birth of an antineutrino (written as $\bar{\nu}$) replaces the death of a neutrino:

$$n \rightarrow p + e^- + \bar{\nu}.$$

This reaction is observed. It is the archetype of all beta decays, the neutron being the simplest radioactive nucleus. On the average it takes 10 minutes and releases about 750,000 electron volts of energy. A gamma emission with the same energy available would take place in some 10^{-16} second. That relative slowness is the clear mark of a neutrino reaction.

If we continue to follow the rule about particles and antiparticles, the transformation might also involve the absorption of a positron by the neutron rather than the emission of an electron:

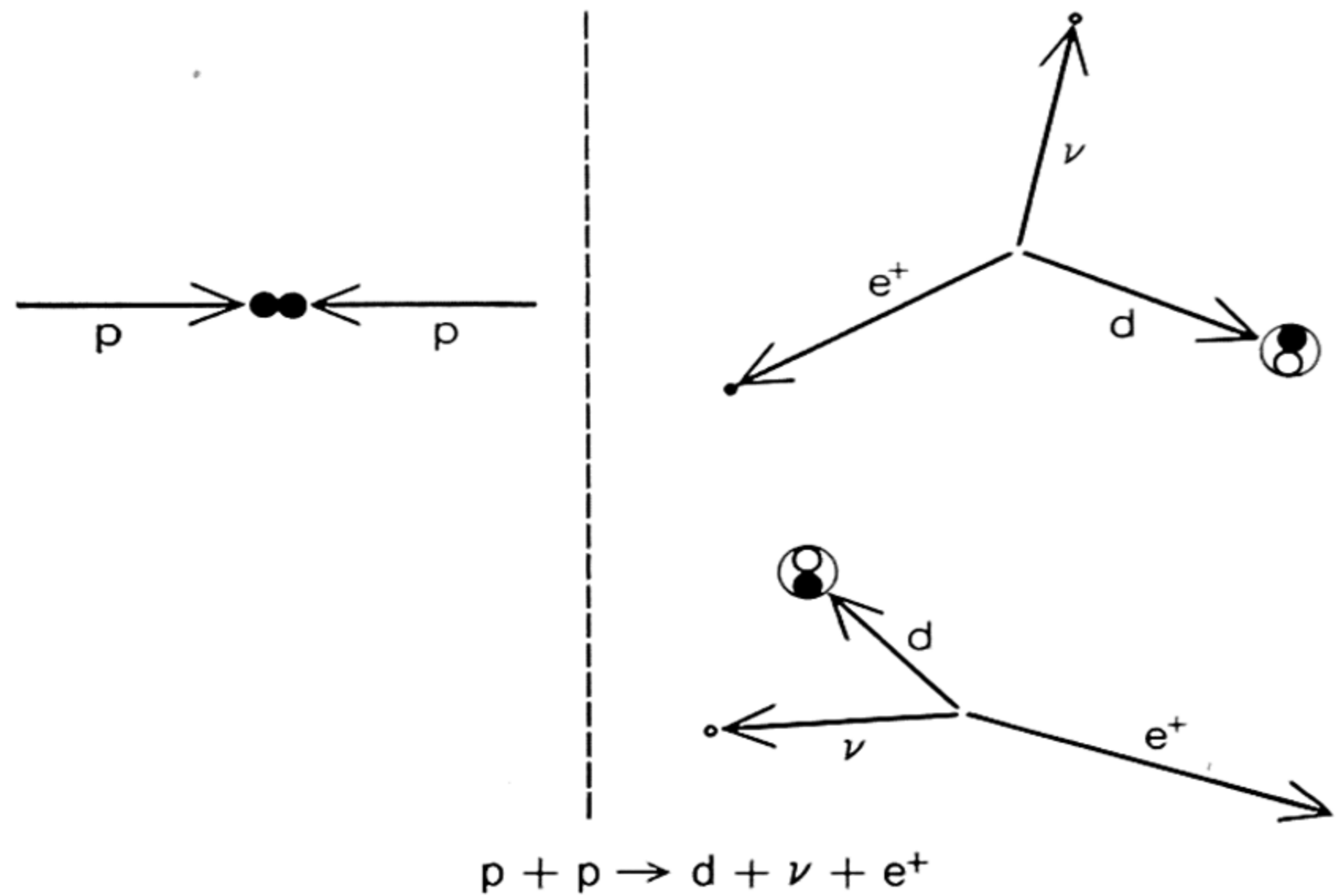
$$n + e^+ \rightarrow p + \bar{\nu}.$$

Moreover, like all particle reactions, this one can be reversed if the energy it releases is supplied from outside. If that were done, we would have a proton absorbing an antineutrino and turning into a neutron and a positron:

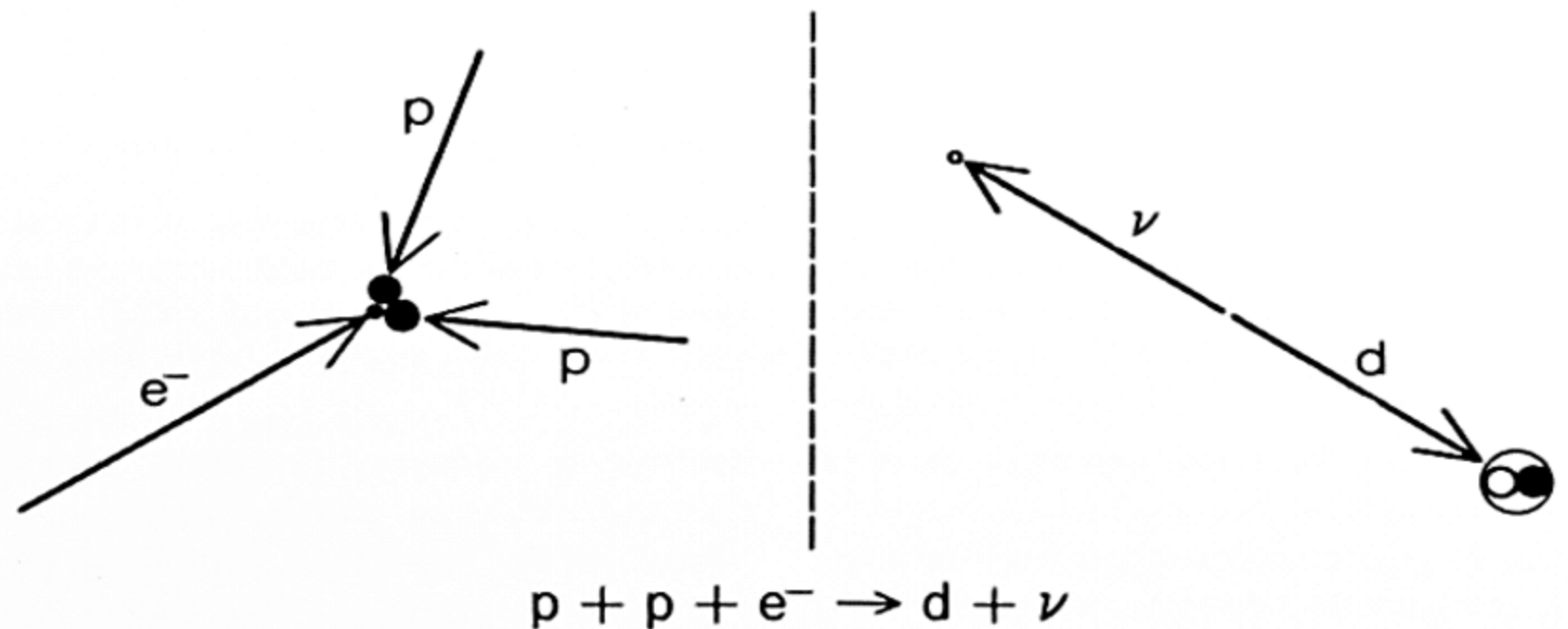
$$p + \bar{\nu} \rightarrow n + e^+.$$

This is the very process that Reines and Cowan observed in their Savannah River experiment. The necessary energy was supplied by the kinetic energy of the neutrinos coming out of the reactor.

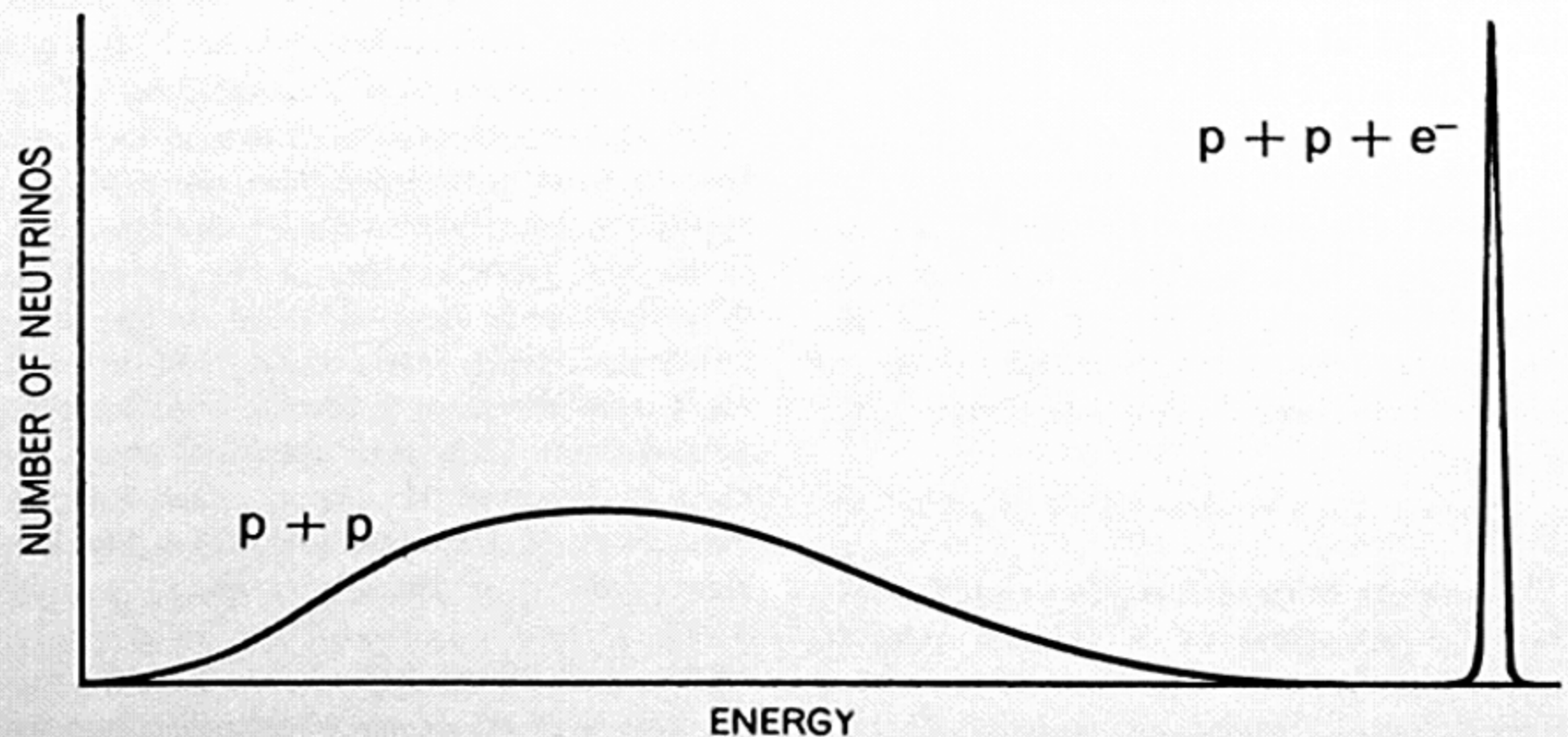
Physicists firmly believe there are many other neutrino reactions that have never been observed, because the theory suggests that any fermion pair can turn into any other. Which of the various processes would supply the material of neutrino astronomy? Armed with the rules of interaction already mentioned, and with more quantitative formulae from which the rates of each reaction can be predicted, some physicists have



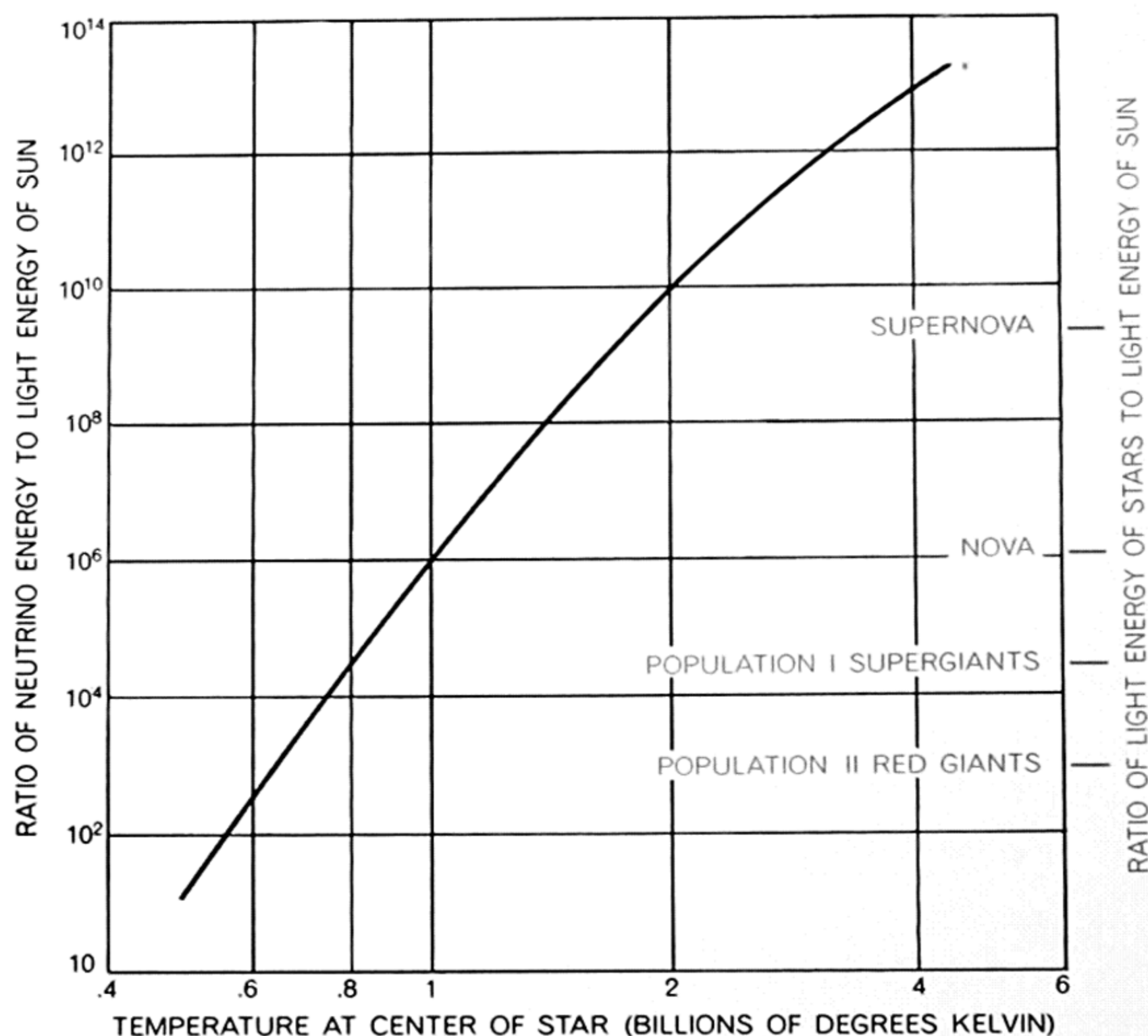
REACTION IN CORE OF SUN, in which two protons (p) come together to form a deuteron (d), also releases a neutrino (ν) and a positron (e^+). Because three particles carry away momentum, the energy and direction of neutrinos vary. Two possible cases are shown.



RELATED REACTION, in which two protons and an electron (e^-) coalesce to form a deuteron and a neutrino, has a smaller but finite probability. With only two product particles, they can move off only in opposite directions, so that neutrinos have constant energy.



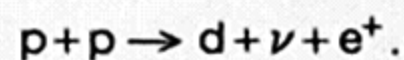
ENERGY SPECTRUM OF NEUTRINOS from the sun has two parts: a continuous band from the reaction shown at top of page and a line from the reaction shown in middle. The line has some width because of thermal motion of the electrons participating in the reaction. This chart is based on the work of Hubert Reeves of the University of Montreal.



NEUTRINO "LUMINOSITY" of stars is plotted against the temperature at their centers. Luminosity is expressed as the ratio of energy emitted by stars in form of neutrinos to the light energy emitted by the sun. Some optical luminosities are shown at right. This chart is based on the work of Hong-Yee Chiu of the National Aeronautics and Space Administration.

set out to identify the neutrino sources in the stars.

The long-lived thermonuclear furnace that is our sun derives its heat from a neutrino-emitting reaction. For some years it has been clear that the combination of two protons to yield the lighter deuteron (a compound of one proton and one neutron) is the first step in the release of solar energy. The reaction is



Since *d* (the deuteron) is simply a close association of *n* and *p*, what has happened from the point of view of the elementary fermions is simply that one proton has turned into a neutron:



The energy release is made possible only by the presence of a second proton. Occasionally two such protons, which cannot form a stable nucleus at all, collide and remain close together for a tiny fraction of a second. If during that time a neutrino is emitted, one proton has become a neutron, and the neutron and proton, still close together, can bind into

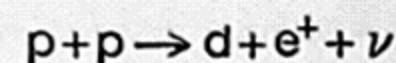
the stable deuteron nucleus, releasing the needed energy. The process takes place, so to speak, *en passant*. And yet so numerous are the collisions at the 15-million-degree temperature and the density (10 times that of lead) of the hot hydrogen gas of the sun's core that this infrequent neutrino emission makes the sun glow. An average proton makes more than 10¹⁶ collisions each second under central solar conditions. Only once in a few thousand years will it have the luck to penetrate the electrostatic repulsive barrier so deeply that the collision will permit a try at the formation of a deuteron. And of these lucky close collisions, only one in 10 million will emit a neutrino and form a deuteron in actual fact. It is the neutrino emission that determines the time scale for the evolution of the sun and thus for the development of life on earth.

The turbulent, hot, dense core of the sun is, then, a source of neutrinos. In the first step of the nuclear reactions there—the formation of a deuteron—about half the energy leaves the sun at once in the straight flight of a neutrino. Subsequent steps rapidly burn the deu-

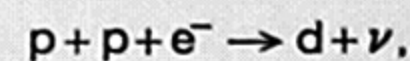
terons to helium without any neutrino emission. The over-all part of the nuclear energy spent in the form of the impalpable neutrinos is about 10 per cent. Even this tithe represents great power, and it would delineate the core of the sun if we could only "see" the neutrinos.

The light of the sun is also born in the core, but it reaches us in quite a different way than do the neutrinos. The difference underscores some of the reasons for dreaming of a neutrino astronomy. Light originates as high-energy X-ray photons deep in the solar interior. Ricocheting again and again, it works its way out to the surface through the enormous bulk of the sun, taking about a million years. In the process each original photon has given birth to a couple of thousand progeny, all of which share the original energy. Neutrinos, on the other hand, interact so little that although they are born in the very core of the star they pass out as though the rest of the great mass were not present. In a neutrino telescope the sun would appear not as a disk but as a tiny hot spot, less than a hundredth the diameter of the sun we see, but emitting about a fourteenth of its power.

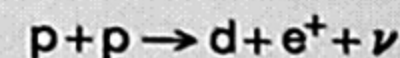
Hubert Reeves of the University of Montreal has given a still stronger reason for trying to build a neutrino telescope. He has reminded us that the reaction



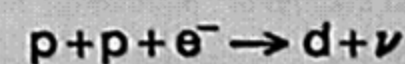
must always be accompanied by the less probable reaction



in which the protons absorb an electron instead of emitting a positron. This will take place only once in a couple of hundred of the standard neutrino emissions. But there is a remarkable difference. In each case the energy is shared among the final particles. The process



produces three particles to share the energy. The motion of any two, say *e*⁺ and *ν*, is balanced by the motion of the third: the recoiling deuteron [see top illustration on page 719]. Therefore direction and energy of the output particles, including the neutrino, are variable. But in the case



the two final particles, *d* and *ν*, can only have equal and opposing momentum.

There is no angle to vary, so the emitted neutrino always carries off the same momentum and hence the same energy. This means that the neutrino energy spectrum is a line spectrum and not a continuous one. The neutrino spectrum from the sun therefore contains both the continuous spectrum of neutrinos from $p + p$ and the distinct line from $p + p + e^-$ [see bottom illustration on page 719]. The line has a small energy width that depends on thermal motion of the particles, and its height depends on the relative probability of the reaction producing it and therefore on the density of the gas where the reaction occurs. To observe the neutrino spectrum would be to determine experimentally the central solar conditions and to fix exactly which reactions are going on. The neutrinos provide the only possible experimental substitute for the "analytical boring machine" of which the British astrophysicist A. S. Eddington wrote, the only way to penetrate the massive shield of a star's body and see what the center is like. It is tantalizing that this message comes constantly to us riding a beam as bright, in order of magnitude, as sunlight itself, and yet we cannot detect it at all!

Evidently neutrino stellar astronomy, if it ever arrives, will bring from the cores of stars experimental information we can now obtain only by analysis of the emitted starlight, many generations removed from the source of energy. Here is an experimental challenge if there ever was one.

It is nearly 25 years since the role of beta-decay neutrinos has been understood in stellar energy release. On that basis alone the emission of neutrinos was a somewhat strange accessory to starlight, but only an accessory. To produce radioactive nuclei that could emit neutrinos nuclear processes were required, which themselves always emitted more energy in particle or photon form.

Some of the other neutrino reactions—recognized in the more powerful theory of recent years—are remarkably different. The simplest and most important process of this still unseen but firmly predicted type turns an electron-positron pair of fermions into a neutrino-antineutrino pair, at a rate easily calculated from the theory of the fermion transformations that are observed. The equation reads

$$e^- + e^+ \rightarrow \nu + \bar{\nu}.$$

All the conservation laws are fulfilled nicely. But if one were to look for this

process in the laboratory he would notice an electron pair disappear with no detectable progeny once in about 10^{20} trials. No one will be that patient.

In such an emission of a neutrino-antineutrino pair no nucleus whatever is involved. The emission does require electron-positron pairs, which might seem an unusual constituent of stellar matter. In fact, however, the pairs are not rare. At high enough temperatures, beginning, say, with a few hundred million degrees centigrade, any volume whatever, even a good vacuum, is populated spontaneously with electrons and positrons. They come from the temperature radiation itself. The photons of this radiation, once they have energy enough, collide with one another to produce electron-positron pairs. Of course the pairs will quickly annihilate back to radiation again. The pair of equations that represent this sort of radiative chemistry are

$$\gamma + \gamma \rightarrow e^+ + e^-;$$

$$e^+ + e^- \rightarrow \gamma + \gamma.$$

Here the Greek letter γ (gamma) represents a high-energy photon. At thermal equilibrium one reaction takes place just as often as the other. The short lifetime of the electron pairs is a guarantee that there are not many present at any time, but they are constantly made and destroyed by collisions in obedience to the general laws of equilibrium. This would have very little importance for the behavior of hot matter and radiation if it were not for the neutrinos. What can it matter if such pairs are transiently formed in the deep stellar interior? Their presence is a mere detail.

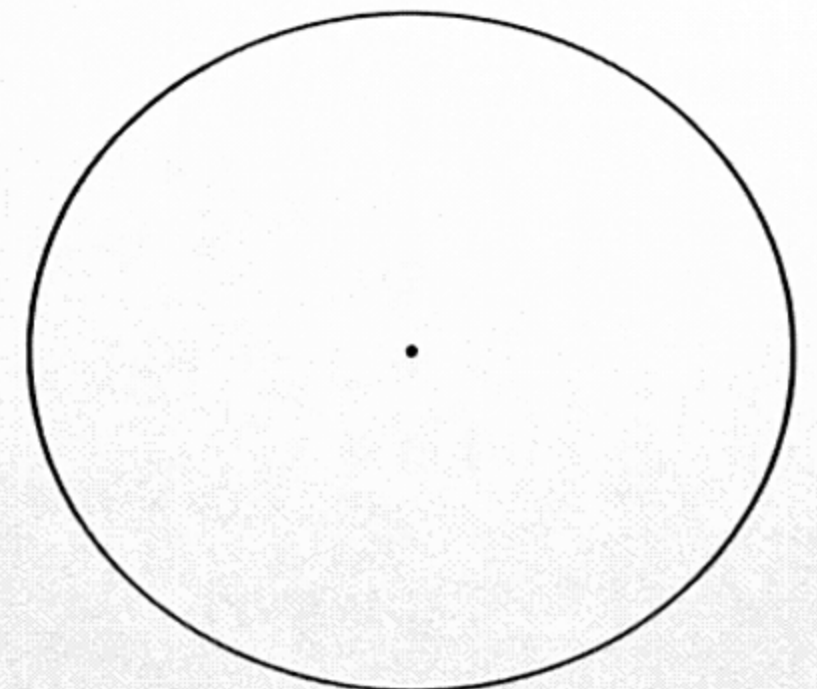
The neutrino emission changes all this. Once in a great many collisions an electron pair will *not* annihilate back into the pair of gamma rays that made it but rather into a neutrino-antineutrino pair. Those particles will never again collide. They escape the region even if its walls are a star, and they seek the depths of space, robbing the hot region of their energy forever. The possibility of neutrino formation means that there is no way to retain thermal energy intact once high enough temperatures are reached to make this reaction important. There are related neutrino-photon processes too, important for somewhat lower and for still higher temperatures. Such energy loss is a newly recognized but fully general feature that the neutrinos give to hot matter of arbitrary composition; the matter loses heat at a definite and by no means negligible

rate. Any walled box at these extreme temperatures must lose all its thermal energy in a matter of minutes by neutrino loss.

In the latter stages of stellar evolution—those stages in which elements of medium weight are formed—temperatures suitable for the emission of neutrinos are expected. The stages are short-lived because the nuclear reactions that maintain the supply of energy are rapid, unlike the slow reactions at the coolish center of modest stars like the sun. Now we realize that they must be still more rapid than had been thought, because they must furnish not only the energy sent out as light but also that sent off in invisible neutrinos. The neutrino luminosity far exceeds the photon luminosity as the temperature rises.

Hong-Yee Chiu of the Institute for Space Studies of the National Aeronautics and Space Administration has shown that at a central temperature of a few hundred million degrees the energy shed by a star becomes dominantly neutrino energy. The hotter the star, the more it is a neutrino star and not merely a visible star. In the last centuries of their lives those few very hot stars headed for explosive death—presupernovae—seem likely to emit billions of times more power in the imperceptible neutrino than they send out as light. Indeed, it appears that the neutrino loss would of itself account for the pre-explosion collapse, although before the collapse can reach really explosive speeds other nuclear events regain control of the fate of hot compressed matter.

It is to be noted that the center of a massive star does not cool when it emits energy but rather collapses and heats. The tendency to cooling implied by energy loss is overbalanced by the



"NEUTRINO SUN" would appear as a small, intense spot (dot) about one-hundredth the diameter of visible sun (circle).

gravitational pull, which causes the hot matter to contract whenever it loses energy. The contraction releases enough gravitational potential energy to supply the lost energy radiated—as light and as neutrinos—and a surplus to heat the compressed star-stuff still more. The heating increases the rate of energy loss by neutrino emission, so the collapse is faster and the loss still greater. Evidently the material has become unstable and is headed toward calamity, of which supernovae are one sign. It is quite possible that we still have not grasped the full meaning of supernovae collapse, which may go well beyond the formation of the most stable nucleus (iron) to produce strange new regions in which not matter but gravitational energy contributes most to the energy account.

We can now see that neutrinos will take away nearly all stellar energy released in the rare but important hotter stages of star evolution. Light removes the energy made available by the burning of hydrogen to helium, and that made by burning helium to carbon as the elements evolve. Neutrinos do almost all the rest, at least until supernova collapse. At that point we are unsure; it is not to be excluded that in gravitational collapse energies as great as all the energy of familiar matter are thrown off in still undetected neutrinos. We can surely assign at least a quarter of all the energy released by stars to neutrinos, and perhaps even the dominant part. The stars that became the supernovae of Tycho Brahe and his student Johannes Kepler, and the star that the Chinese astronomers of the Sung dynasty saw by day, had been invisibly “brightening” in the imperceptible glow of neutrinos for cen-

turies before they finally grew bright for a few months in visible light when the slowly diffusing, much scattered photons finally worked their way to the surface. Neutrino detectors would make possible a supernova-early-warning service.

The end of the neutrino story is still uncertain. It may be that this newly discovered means of removing energy works even in the common smaller and cooler stars, allowing them to cool to the white-dwarf stage without ever reaching giantism and heavy-element-burning temperatures. It may be (as John A. Wheeler and his associates at Princeton University and Bruno M. Pontecorvo, I. B. Zeldovitch and Y. Smorodinsky in the U.S.S.R. have conjectured) that there is as much energy in the impalpable neutrino flux as in matter itself, placed there either by some supercollapse or by strange processes that go back to a possible explosive creation of all matter. It is even possible to see in the neutrino a relation with the very expansion of the universe, an expansion perhaps caused by the pull of an enormous total mass of very low-energy neutrinos. None of these matters is settled, and presumably none can be until the natural stream of neutrinos is detected. Then we might be able to reach some cosmological decisions in the nuclear laboratory.

At the present moment Reines, now at the Case Institute of Technology, and Cowan, now at Catholic University, are mounting schemes that at least point toward a neutrino astronomy. They are setting out to detect neutrinos produced not by stars but by cosmic rays and that have substantially higher energies than those treated in this article. The higher

the energy of the neutrinos, the greater their probability of interacting with other matter.

Already Reines and his student Charles C. Giamati have been able to set a new limit on the higher energy cosmic ray neutrino flux that is close to the rate actually expected. The big detector with which this was accomplished was located 2,000 feet underground in a mine of the Morton Salt Company near Cleveland [see illustration on page 716]. The next step, now in the design stage, is a counter containing 1,000 or more tons of the purest and most limpid water, set deep underground to shield out electrically charged cosmic ray particles and scanned by large photomultiplier tubes. They seek to detect the light radiated by fast particles secondary to the high-energy neutrinos entering the counter.

To be sure, even this heroic effort will fall short of revealing the still more elusive, star-born, electron-associated neutrinos. The experimenters are set too hard a task by the weakness of the neutrino interaction at low energy, the very weakness that allows the particles to signal us so directly from the center of the stars. But perhaps a way can be found. Maybe one day we can use as detectors not the nearly stationary protons and electrons of matter but fast-moving particles, made highly energetic in our own accelerating machines. Someday we may watch for events occurring from collisions of our fast particles against invisible partners, the neutrinos from the depths of space.

Suffice it to say for now that the message is there, bearing information on the cores of the stars and perhaps on the very bounds of space and time.

The Author

PHILIP MORRISON is professor of physics and nuclear studies at Cornell University. He obtained his Ph.D. under J. Robert Oppenheimer at the University of California in 1940 and then worked on the Manhattan Project, first in Chicago and later at Los Alamos. His present research interests are on the borderline between astronomy and particle physics, including the subject matter of his article in this issue. "Neutrino Astronomy" is the eighth SCIENTIFIC AMERICAN article of which he has been the author or coauthor.

Bibliography

EMISSION OF PHOTONEUTRINOS AND
PAIR ANNIHILATION NEUTRINOS FROM

STARS. Hong-Yee Chiu and Robert C. Stabler in *The Physical Review*, Vol. 122, No. 4, pages 1317-1322; May 15, 1961.

THE NEUTRINO. Philip Morrison in *Scientific American*, Vol. 194, No. 1, pages 58-68; January, 1956.

THE NEUTRINO. C. S. Wu in *Theoretical Physics in the Twentieth Century*, edited by M. Fierz and V. F. Weisskopf, pages 249-303. Interscience Publishers, Inc., 1960.

NEUTRINO INTERACTIONS. Frederick Reines in *Annual Review of Nuclear Science*, Vol. 10, pages 1-26; 1960.

THEORY OF THE FERMI INTERACTION. R. P. Feynman and M. Gell-Mann in *The Physical Review*, Vol. 109, No. 1, pages 193-198; January 1, 1958.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

SEMICONDUCTOR PARTICLE-DETECTORS

by Olexa-Myron Bilaniuk

This new type of counter is supplanting all others in the field of low-energy nuclear physics. Among its advantages are rapid response and the ability to distinguish particles close in energy.

Nuclear physics became possible when experimenters first learned how to detect the particles emitted by radioactive materials. Today the study of nuclear structure is still largely a matter of counting and measuring the fragments that excited atomic nuclei eject to rid themselves of excess energy. Recently the measurements have been made both easier to perform and more precise through the application of semiconductors. Just as the transistor has all but pushed the vacuum tube out of the electronics shop, so are semiconductor detectors fast supplanting bulkier and less efficient counters in the nuclear laboratory.

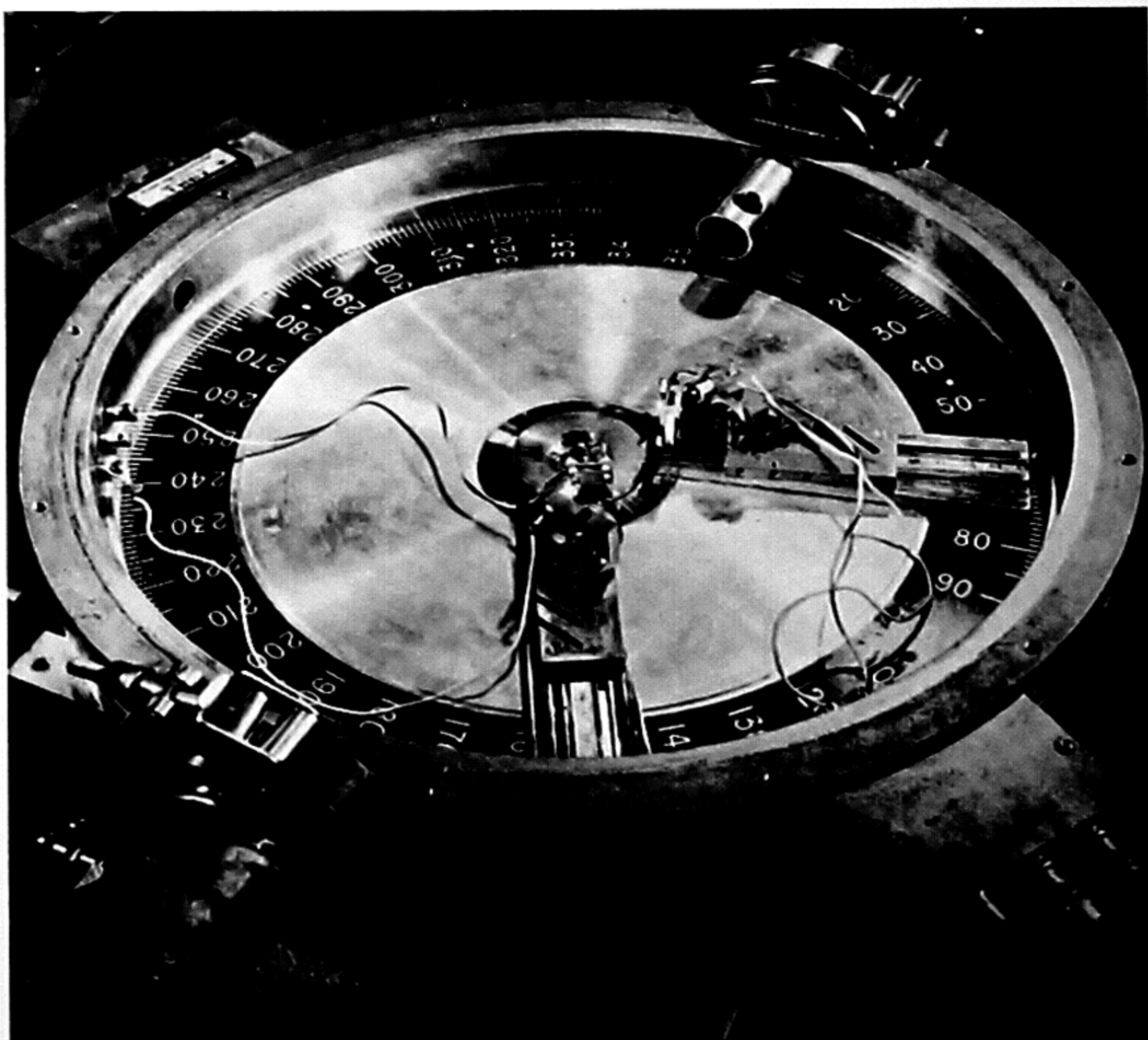
By current multibillion-electron-volt standards the particle energies in nuclear radiation are modest indeed: they are measured in millions of electron volts (Mev) or, at most, a few tens of Mev. As it happens, it is precisely in this low-energy region that semiconductor counters are best suited to function. Not only can they detect low-energy particles efficiently, they can also differentiate between tiny differences in energy.

The classical detector in low-energy physics has been the ionization chamber. A brief consideration of its operation will help to make clear the advantages offered by the new devices. The chamber is filled with an inert gas and contains a pair of electrodes, which are maintained at a potential difference of about 1,000 volts. A nuclear fragment entering the chamber through a thin window ionizes some of the gas atoms. The electrons and positive ions are swept out of the gas and onto the electrodes, giving rise to an electric pulse proportional in amplitude to the number of ion pairs produced. This in turn is proportional to the energy lost by the fragment in the chamber. The duration of the pulse depends on the time required to

sweep up the liberated charge. (In contrast to its well-known relative, the Geiger counter, the ionization chamber does not employ a strong electric field to multiply the charges produced by the incoming particle. In order to measure energy, the chamber must respond with pulses of different sizes to different quantities of liberated charge.)

The quality of the counter is deter-

mined principally by its ability to distinguish between fragments of nearly equal energy; or, more strictly, between fragments losing nearly equal quantities of energy in the chamber. Because many nuclear measurements require a distinction among almost simultaneous events, a counter should produce brief signals to achieve good time resolution. Brief signals provide the additional bonus of



SEMICONDUCTOR COUNTERS are photographed in an experiment on neutron-neutron interaction at the Brookhaven National Laboratory. View at left shows vacuum chamber in which a target material, placed at the center, is bombarded by cyclotron beam emerging from tube at top. Counters are mounted on two arms. In close-up view at right the circular

short recovery time, enabling the counter to respond to a rapid succession of particles.

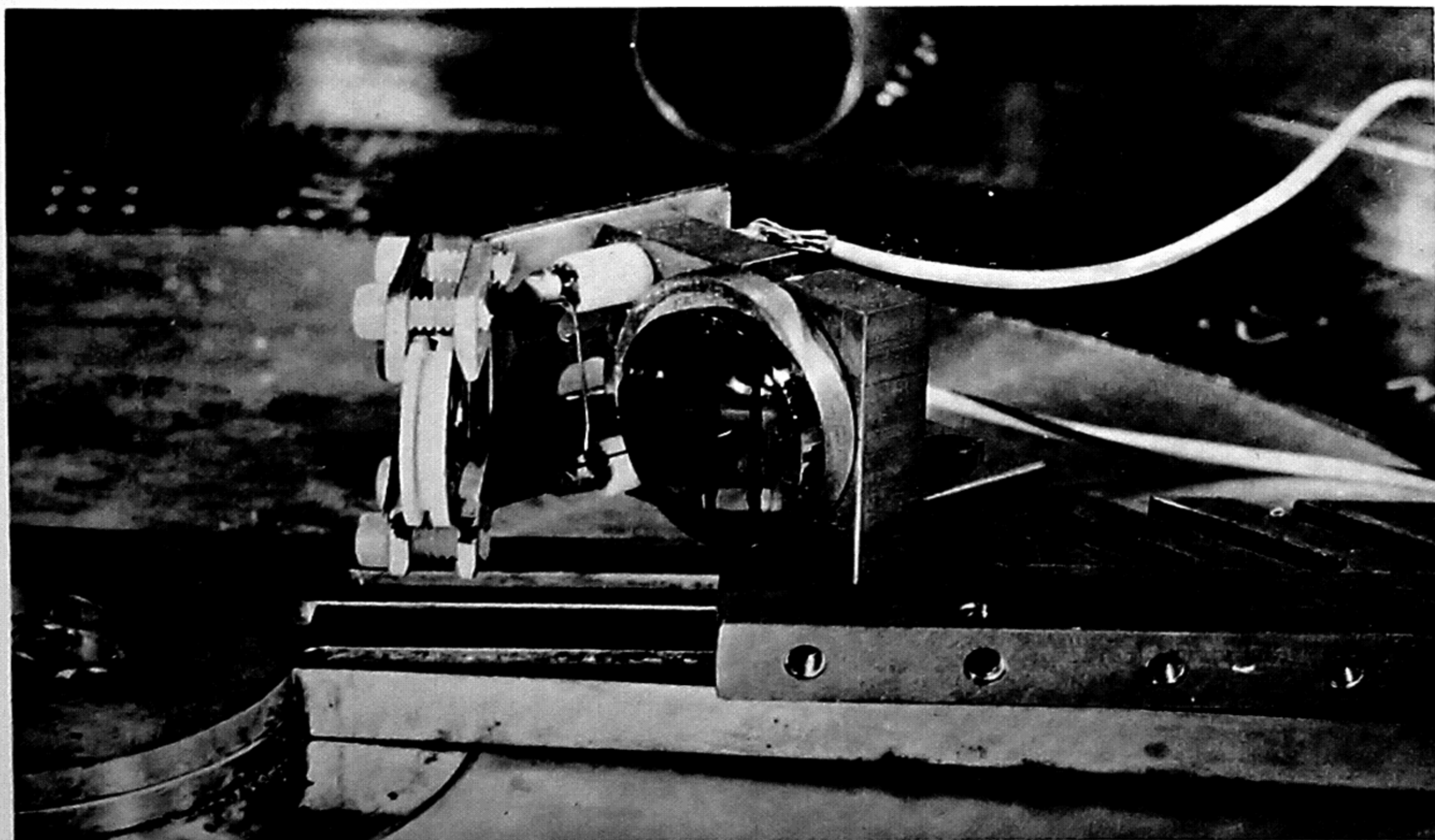
Both the energy resolution and the time resolution of an ionization chamber are easy to estimate. The theoretical limit for energy resolution is set by the fact that it takes, on the average, 27 electron volts of energy to ionize one gas atom. If a nuclear fragment loses one Mev in the gas, it produces some $1,000,000 \div 27$, or 37,000 ions. This average figure is subject to an uncertainty arising from the chance element in the process of collision and ionization. On the assumption of a "Poisson distribution," or a bell-shaped probability curve, the uncertainty is measured by the square root of 37,000, or 192. A detailed analysis shows that the curve is not exactly bell-shaped and that the uncertainty is somewhat smaller, but the simpler computation correctly reflects the dependence of the uncertainty on the amount of energy required to produce an ion pair. In the present example this uncertainty sets a theoretical limit of $192 \div 37,000$, or .52 per cent, on the energy resolution of the ionization chamber.

Time resolution is simply a question of the speed with which the electric field clears out at least the majority of the liberated charges. It is determined primarily by the distance the charges have to travel and by the intensity of the electric field. In a practical ionization chamber it is hard to resolve pulses separated by less than 10 microseconds (millionths of a second). In terms of counting rate this means that if the number of accidental coincidences is to be less than 1 per cent of the registered counts, the rate must be limited to fewer than 1,000 counts per second.

An ionization chamber must be fairly large because to lose a measurable amount of its energy a nuclear fragment must cover a considerable distance in a gas. Here the advantage of a solid-state device is most clearly apparent. The great stopping power of scintillation crystals has made them the most popular detectors for low-energy nuclear research [see "Scintillation Counters," by George B. Collins; *SCIENTIFIC AMERICAN*, November, 1953]. Their energy resolution is limited, however, to about

3 per cent at best, and they are inseparable from the intricate photomultiplier tube. On both counts the semiconductor counter is superior, and it is displacing not only the gas-filled ionization chamber but also its own solid-state kin, the scintillation crystal.

In their physical appearance semiconductor counters resemble scintillation crystals, but in operation they are more closely related to ionization chambers. Charges created by incoming particles are collected by electrodes at the surface and converted to electric pulses. The first devices to be designed on this principle did not make use of the peculiar properties of semiconductors but were made simply of nonmetallic crystals, such as diamond or silicon. The carbon atoms in a diamond crystal are bound to one another in such a way as to tie all the atomic electrons to their respective lattice points. No charges are available to carry a current even if a high electric field is applied across the crystal. In solid-state parlance we say that the valence band, where the electrons are tied to their lattice sites, is separated by a wide, forbidden energy



disk facing 45 degrees from the axis of the arm is an "n-p junction" counter (see text), which counts the particles from target and measures their energy. The rectangles on one-inch silicon disk outline two separate counting areas. To the left of disk, seen edge on,

is a "surface barrier" counter, which measures the rate of energy loss of the particles. The information from both counters serves to identify particles. The experiment is a joint project of the Bell Telephone Laboratories and Brookhaven National Laboratory.

gap from the conduction band, in which the charges are free to move through the crystal. When a nuclear fragment enters the crystal, it knocks electrons from the valence band into the conduction band, thus creating mobile charge carriers. These consist of both the liberated electrons and the positive "holes" they vacate. An external electric field can now sweep out the charge carriers, giving rise to a pulse of current.

In addition to its much higher stopping power compared with gas, the crystal counter surpasses the ionization chamber in energy resolution and recovery time. To activate one charge pair in the diamond crystal requires only nine electron volts, on the average, as against 27 for argon. Repeating the earlier calculation shows that the loss of one Mev of energy by a fragment activates 111,000 electron-hole pairs, with an uncertainty of the square root of 111,000, or 333. This gives an efficiency limit of .3 per cent. Even more significant is the improvement in time resolution. Principally because much shorter distances are involved, the activated charges are swept out from the crystal in a few hundredths of a microsecond instead of a few microseconds.

In spite of its great advantages, the crystal counter has not been a success. Two drawbacks are responsible: dark current and polarization. Random

thermal motions continually activate charge carriers in the crystal, even in the absence of an ionizing nuclear projectile (that is, even in the "dark"). Under the influence of the external electric field, they give rise to an ever present fluctuating current that constitutes electrical noise above which the useful signals must be detected.

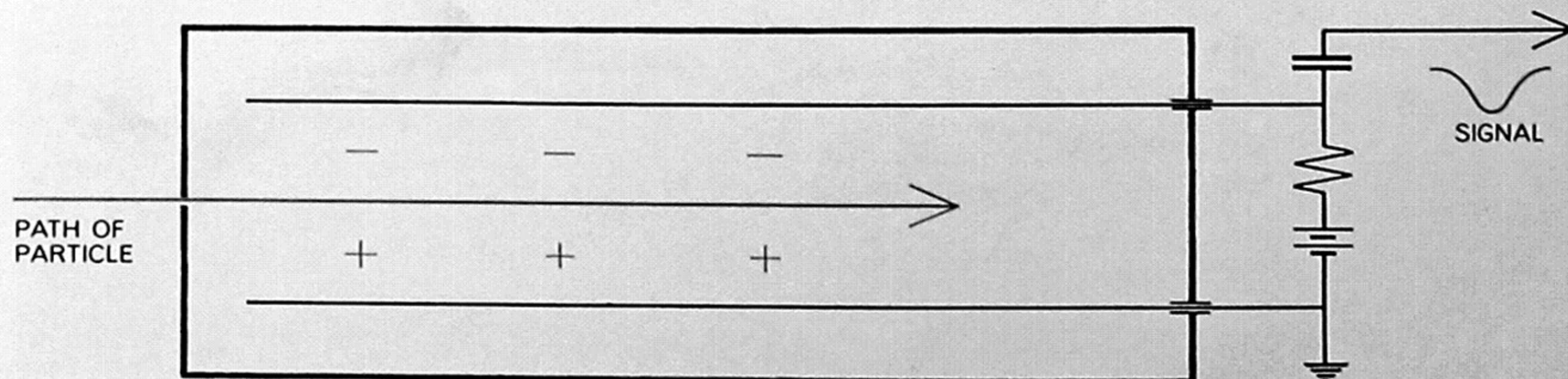
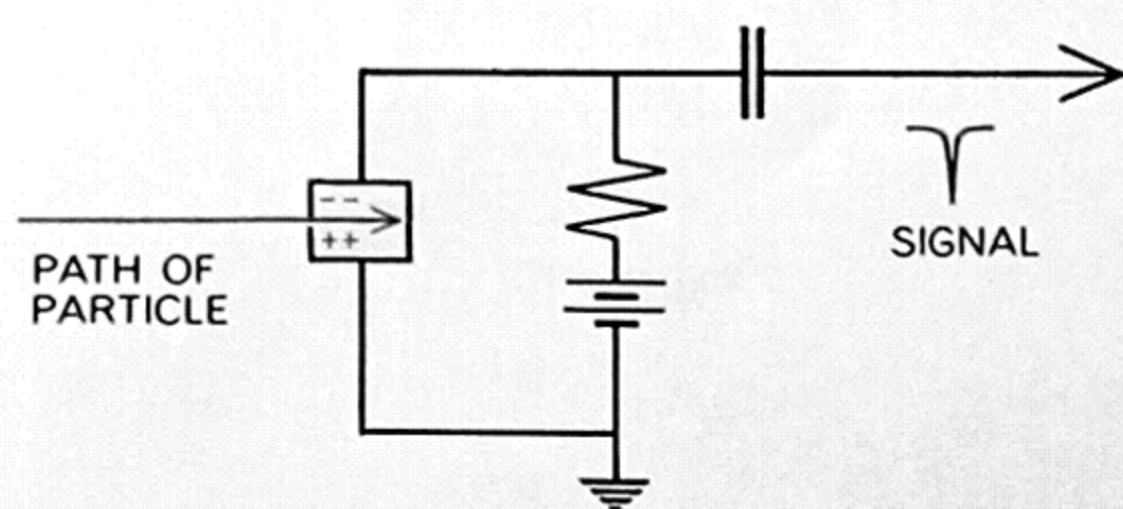
The dark current is insignificant in diamond because of the large width of the forbidden gap and the presence of so-called trapping centers arising from localized crystal imperfections. Depending on the type of imperfection, a trapping center may catch electrons or holes. Both types of center exist in every crystal, and they waylay a substantial fraction of the charge carriers activated by the bombarding particles. These trapped charges build up an electric field inside the crystal, which opposes the externally applied field. Even at moderate counting rates polarization considerably reduces the effective field strength. Not only does this lower the detection efficiency, it also destroys the proportionality between particle energy and collected charge, which is essential for good energy resolution.

By choosing a crystal such as silicon, which is similar to diamond but has a smaller forbidden gap, the effects of polarization can be considerably decreased. The bombarding particles release enough charge carriers to fill the

traps and still produce substantial current. But the gain is offset by a prohibitively large dark current. A slab of high-purity silicon a quarter of a centimeter thick and one square centimeter in area has a resistance of approximately 2,500 ohms. Applying 250 volts across the slab gives rise to a dark current of a tenth of an ampere. Merely to dissipate the resulting 25 watts of heat becomes a problem, to say nothing of the difficulty of detecting the pulses that are caused by radiation.

What is wanted is a crystal with high resistance, to minimize dark current, and with minimum tendency to polarization. A fairly effective compromise that has been developed recently makes use of a silicon crystal judiciously doped with phosphorus and gold. Phosphorus atoms have five outer, or valence, electrons, one more than can fit into the interatomic bonds in the silicon crystal. The extra electrons are easily detached from their atoms to become mobile charge carriers. Because the majority of carriers are then negative, phosphorus-doped silicon is called an n-type semiconductor. The availability of electrons from the donor phosphorus atoms means that n-type silicon has much lower resistance than pure, or "intrinsic," material. Adding the right number of gold atoms just offsets the decrease in resistance by providing deep trapping centers to capture the mobile electrons. The net effect is that a crystal of ordinary silicon achieves the resistance of extremely pure, intrinsic material.

At moderate counting rates detectors made of silicon doped with phosphorus and gold work remarkably well. Moreover, the statistical limit on energy resolution is very low, since the average energy needed to generate a charge pair



CRYSTAL COUNTER (top), which was forerunner of semiconductor counters, is compared schematically with classic ionization chamber (**bottom**). Both are shown in approximately actual size.

In both an incoming particle releases mobile charges, which are swept away by an externally applied voltage and which produce a brief output voltage pulse. Crystal gives much sharper pulses.

in silicon is only 3.5 Mev. In the case of a particle losing one Mev of energy the limit is only .2 per cent. At higher counting rates, however, gold-doped silicon crystals are still prone to polarization. In addition, they must be operated at liquid nitrogen temperatures so that thermal motions do not lift too many charge carriers across the narrow forbidden gap.

The chief stumbling block for the homogeneous crystal counter, at least in the present state of technology, is the apparent impossibility of achieving high resistivity without recourse to trapping. An elegant resolution of this problem has been found in the surface-junction counter.

Consider a silicon crystal in the interior of which one lattice atom in every million million or so is replaced by a boron atom. Boron has three valence electrons—one fewer than are required to satisfy the bonds to all neighboring atoms in the crystal. Therefore every boron atom constitutes an acceptor site, where electrons can easily lodge. Electrons moving into the sites leave behind positive holes that can move through the crystal and carry electric current. For this reason the material is called p-type. Suppose that one surface of the p-type crystal is doped with phosphorus to make it n-type. Then some of the extra donor electrons from the surface will lodge themselves in acceptor sites underneath, creating a narrow polarized region that is positive on one side and negative on the other [see illustration on page 729]. This dipole layer of bound positive and negative charge gives rise to an internal electric field that maintains the "depletion" region free of charge carriers and gives it very high resistivity. An external electric field applied in the direction of the internal field serves to widen the depletion region. In contrast to gold-doped silicon, the junction has no deep trapping centers to cause polarization.

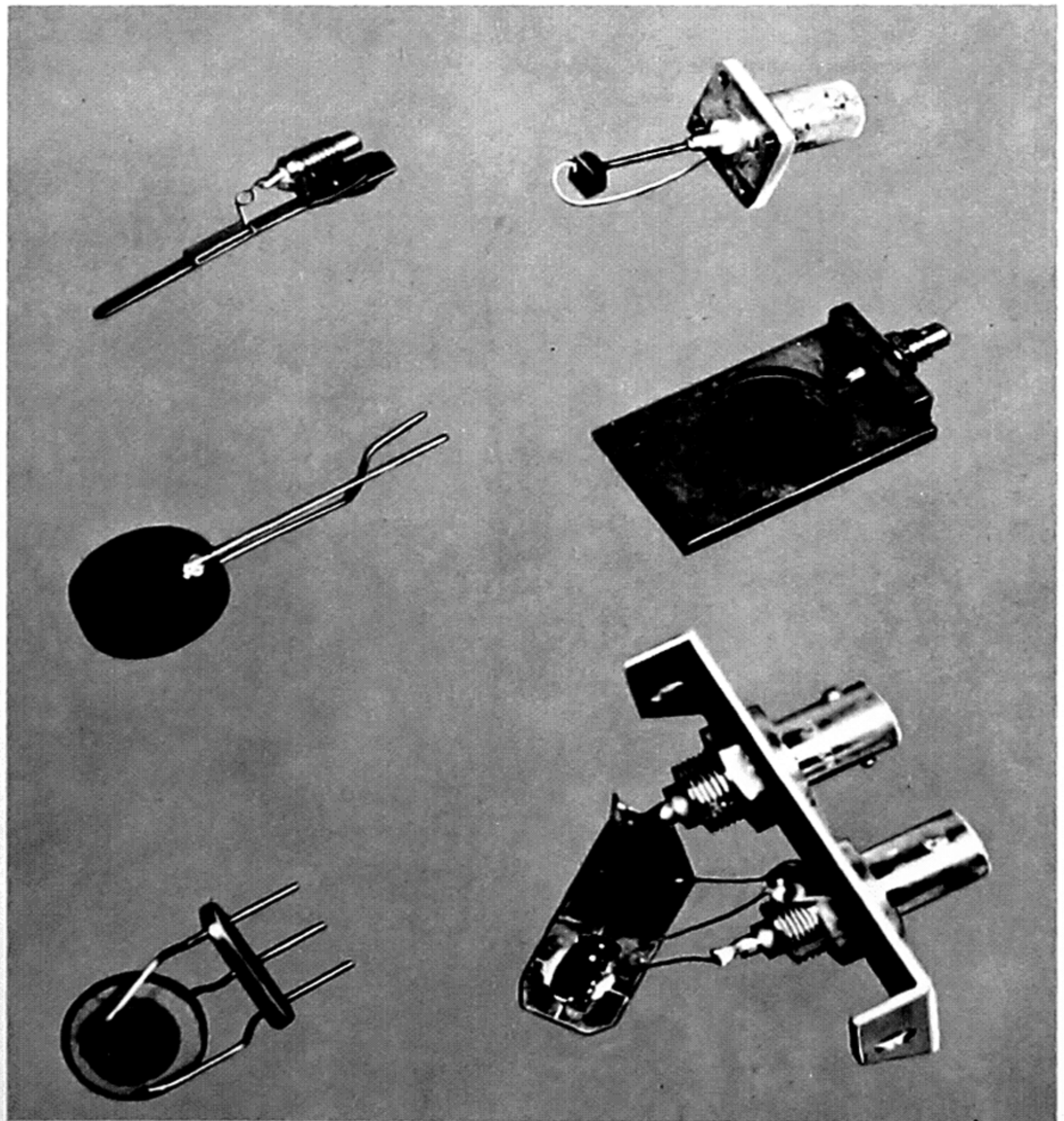
When an ionizing nuclear fragment passes through the junction, it leaves a plasma of conduction electrons and holes in its wake that is swept out by the electric field within a few nanoseconds (thousands of a millionth of a second). This is the fastest detector recovery time known. The energy resolution, determined by the low energy needed for creation of an electron-hole pair in silicon, is still at its record low. (If germanium, with a narrower forbidden gap, is used instead of silicon, the limit to the energy resolution lies even lower.

But germanium junction counters have to be cooled, whereas silicon works perfectly well at room temperature.)

Similar to the surface-junction detector, both in quality and operation, is the silicon (or germanium) surface-barrier detector. It is designed to take advantage of the fact that the undamaged surface of a crystal can play the role of acceptor. Donor electrons from the interior of a phosphorus-doped, n-type silicon crystal collect on the surface, establishing a high-resistivity depletion region near the crystal surface, exactly as in the junction counter. A thin film of gold evaporated onto the surface makes it possible to widen the depletion layer by applying an additional voltage between the surface and the interior of the crystal. An ionizing

particle entering the depletion region gives rise to a current pulse that can be collected at the gold film.

Although surface-junction and surface-barrier counters are unexcelled in their energy resolution and speed of signal, the limited depth of their depletion layer constitutes a serious weakness. To distinguish among particles of different energy the detector must stop all the particles within its sensitive region so that all the energy of each one contributes to the output pulse. The depth of the sensitive depletion layer increases as the square root of the applied voltage times the resistivity (determined by the purity) of the crystal. Even with externally applied fields as high as 400 volts and the purest available silicon, the



VARIOUS TYPES OF COUNTER discussed in the text were built in the author's laboratory at the University of Rochester. Raised rectangle in middle of strip at top left is a surface-barrier counter. Entire assembly is about two inches long. The other devices are: a surface-junction counter (*top right*), an n-i-p counter (*middle left*), a thin transmission counter (*middle right*), a p-n-i-p counter (*bottom left*), in which transmission counter is incorporated in the n-i-p device, and a combined transistor-counter (*bottom right*), which amplifies its own signals. Commercial firms have begun to make some of the devices.

depth of the sensitive region is only one millimeter. Therefore at present depletion-layer counters can be used to detect only particle that lost all their energy before they traverse one millimeter of silicon. Since the rate at which the particles lose energy is proportional to the square of their charge, the shall-

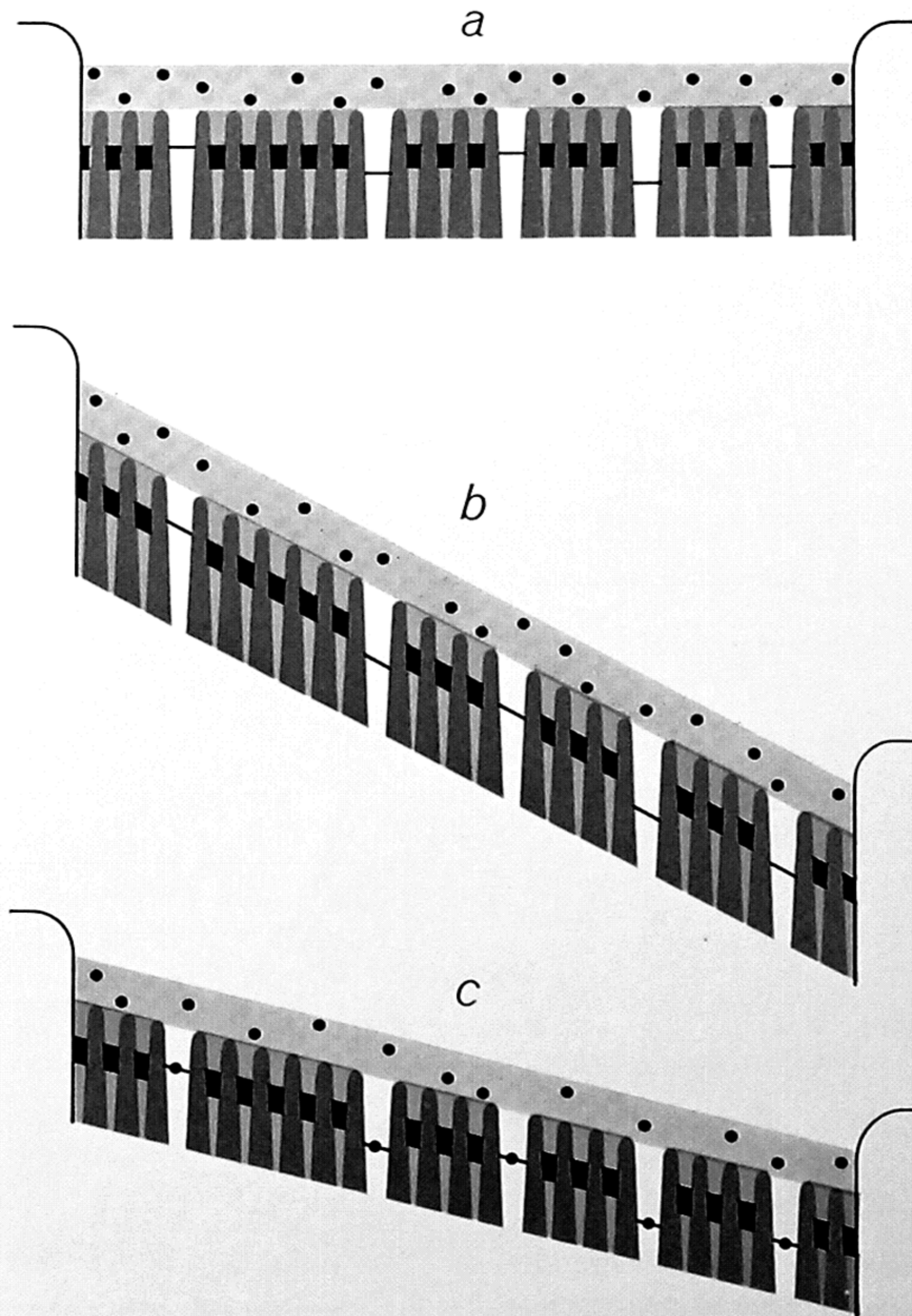
owness of the sensitive region is no handicap in working with multiply charged fragments. In most cases they are completely stopped in silicon after traveling only a fraction of a millimeter. For this reason these counters are admirably suited to the detection of nuclear fission fragments and have recently made

possible fission measurements of unequalled precision.

Even alpha particles and helium 3 ions, with their charge of two units, are easily detected, and their energy measured, within the sensitive region, as long as their energy does not exceed some 40 Mev. Most investigations of nuclear structure involve energies below this value, so that surface-junction and surface-barrier counters are now used almost exclusively for alpha spectroscopy and similar experiments. The trouble begins as soon as protons or deuterons (nuclei containing one proton and one neutron) are involved. The rate of energy loss for protons is only about a quarter of that for alpha particles, so that a detector with a sensitive region wide enough to stop 40-Mev alphas will be good for protons only up to 10 Mev. Tilting the crystal at an angle to the incoming particles extends the range to somewhat higher energies, but the basic limitation remains. Of course this limitation makes depletion-layer counters completely unsuitable for high-energy nuclear physics, where particle energies reach billions of electron volts.

A considerable widening of the sensitive depletion region is achieved in the n-i-p (n-intrinsic-p) version of the junction counter. This device can be envisaged as a cross between an n-p junction and a crystal counter. It was first made in France by diffusing phosphorus and boron into the opposite faces of an ultrapure silicon slab. The result was a crystal with three distinct zones: a thin n-type surface layer, an intrinsic (undoped) interior and a p-type surface layer. As in the case of the n-p junction, some of the donor electrons drift over to occupy some of the acceptor sites, thereby depleting portions of the n and p layers of charge carriers. Moreover, the electric field thereby created bridges the intrinsic region, and so it too is kept free of carriers. Therefore practically the entire crystal becomes sensitive to ionizing rays. The counter works well, but it is difficult to make. Even extremely minute quantities of impurities in the silicon supply so many charge carriers that the central region can no longer be maintained in a depleted state.

An ingenious substitute for ultrapure silicon has been developed at the General Electric Research Laboratory by E. M. Pell. He begins with p-type silicon and diffuses lithium atoms into one face of the crystal. Lithium, with its single outer-shell electron, acts as a donor impurity and creates an n-p surface junction. Unlike the donors mentioned previously, however, lithium atoms do



CONDUCTION IN CRYSTAL can take place only if charge carriers (black dots) are raised from valence energy band (lower gray segments) to conduction band (upper gray band), where they are above the potential energy barriers (dark-colored shapes) separating the atoms of the crystal. Applying an outside voltage provides a potential "hill" (b) down which the carriers can move. Imperfections in the crystal are traps (black lines) for carriers. When the traps are filled (c), an internal voltage partly offsets the applied voltage.

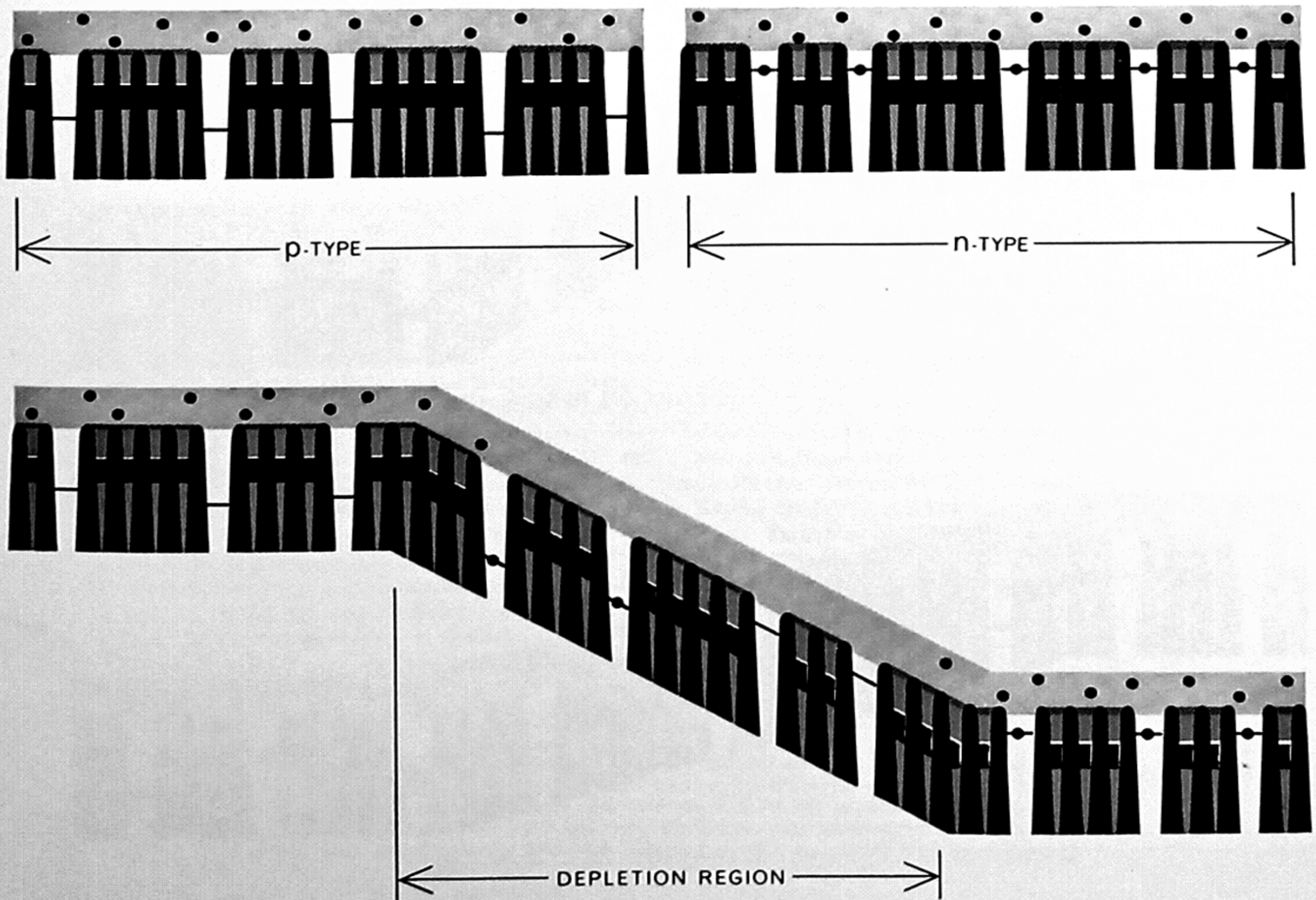
not fit themselves into the crystal lattice proper but remain "interstitial." They are small enough to wander through the lattice. If a voltage is applied to the crystal in the right direction, lithium atoms that have given up their valence electrons will be forced into the p-type interior. There each ion eventually encounters an acceptor site occupied by an electron and hence negatively charged. The positive lithium ion neutralizes, or "compensates," the acceptor ion. In the process it becomes neutral itself and is no longer acted on by the outside voltage. The result is that all the acceptor sites within a rather wide zone are eventually compensated, and the material here exhibits properties of intrinsic silicon of very high resistivity. In effect the crystal now has an n-i-p configuration. Many laboratories today are fabricating n-i-p counters by the lithium drift method. Sensitive depths up to two millimeters are achieved more or less

routinely. A few experimenters have made high-quality n-i-p counters twice as thick, and Jack H. Elliott at the Radiation Laboratory of the University of California has succeeded in extending the compensation to a depth of six millimeters, which appears to be a record. Although the manufacture of thick semiconductor counters is still largely an occult art, it will undoubtedly be reduced soon to a standardized procedure.

Interestingly enough, at the same time that physicists are straining to make thicker semiconductor counters, they are also trying to produce extremely thin ones. The latter are useful in distinguishing different types of particles of the same energy. Suppose that 20-Mev alpha particles are to be counted in the presence of 20-Mev deuterons. Alpha particles at this energy have a range of .2 millimeter in silicon; the deuterons penetrate to 1.5 millimeters. A counter

with a depletion region only .2 millimeter thick will completely stop the 20-Mev alpha particles within its sensitive region, while absorbing less than two Mev of energy from the deuterons. Consequently the alpha signals are about 10 times larger than deuteron signals and can easily be selected electronically. It should be noted that although the experiment requires a narrow sensitive zone, the crystal need not be thin.

Now consider the problem of counting the deuterons rather than the alphas in the mixed beam. As long as the 20-Mev deuterons do not have to be further distinguished from, say, 19.7-Mev deuterons, the solution is straightforward. The alphas can be absorbed away by covering the counter with a thin foil. The more penetrating deuterons pass through the foil, leaving only two Mev of energy behind, and can then be detected without interference. But in traversing the foil deuterons undergo a



P-N JUNCTION COUNTER is made, in effect, by joining p-type and n-type semiconductors. P-type contains empty "acceptor sites" (black lines); n-type contains "donor sites," each with an easily

lost electron (black dot). At junction electrons from donor sites drift into acceptor sites, producing a voltage that keeps the "depletion" region (slanted section) free of mobile carriers.

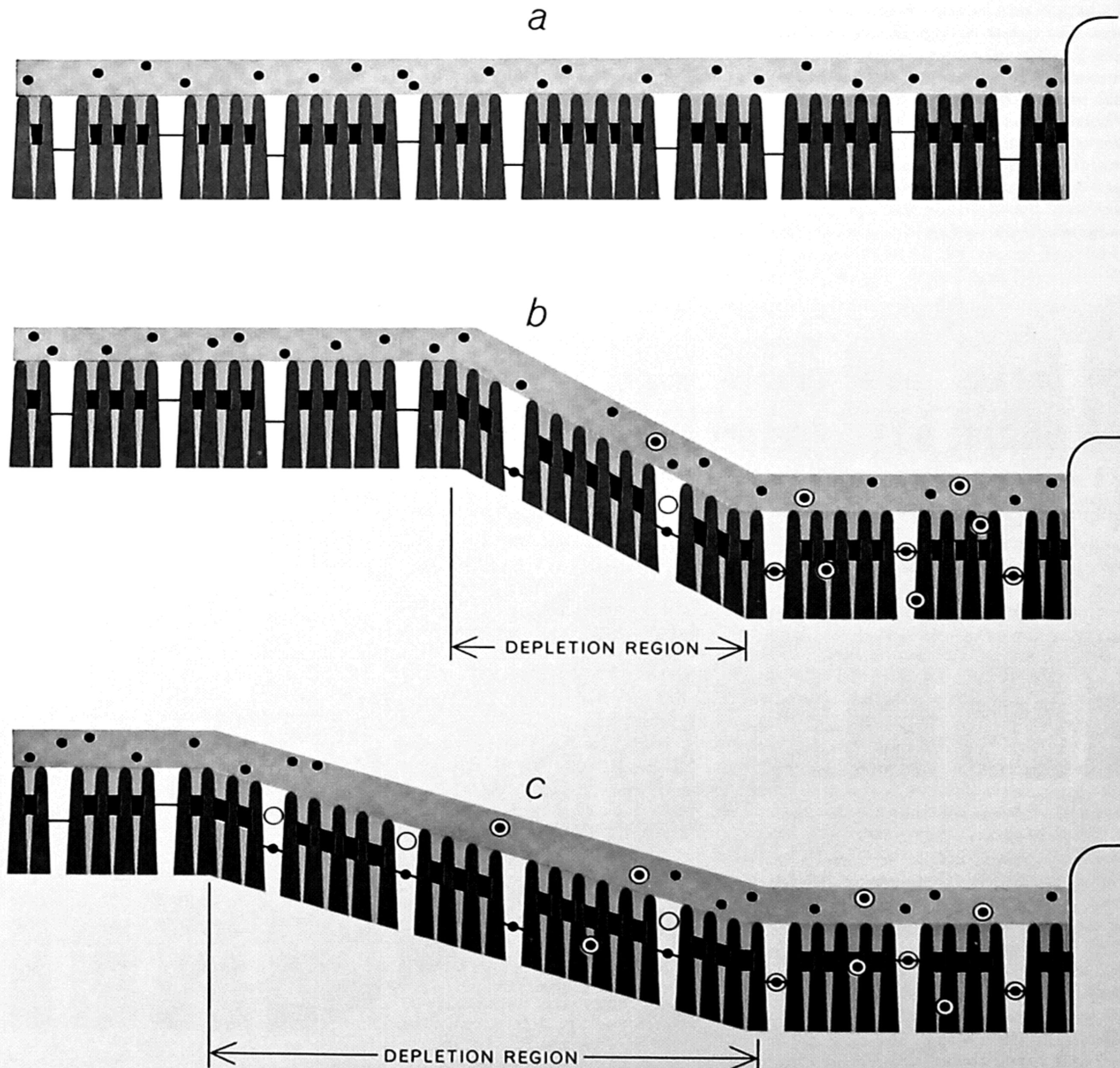
process called straggling, which spreads their energies. If the 20-Mev deuterons are accompanied by 19.7-Mev deuterons, the two groups become indistinguishable.

This is where the extremely thin transmission counter comes in. Harvey E. Wegner of the Los Alamos Scientific Laboratory has demonstrated that in traversing a thin crystal different par-

ticles give rise to pulses proportional to the rates at which they lose their energy. Therefore alphas produce much larger pulses than deuterons. The smaller deuteron signals from a thin sampling counter can be used to "gate" a second, ordinary counter so that it will respond only to the deuteron and ignore all the pulses arising from the alphas. Since the sampling counter is very thin, straggling

is not significant and the high resolution of the main counter can be fully realized.

Making extremely thin transmission counters involves as much magic as does fabricating thick depletion layers. The main bugbear is the fragility of thin silicon wafers. Workers at the Bell Telephone Laboratories have recently produced a transmission counter a hundredth of a millimeter thick—a remark-



N-I-P COUNTER is made by diffusing lithium atoms (*open circles*) into one surface of p-type semiconductor. Lithium can donate its single outer electron (*black dot in circle*) to acceptor sites, converting surface to n-type (*"b" right*) and forming an n-p junction

depletion region (*"b" center*). Lithium ions pushed into crystal by an outside voltage "compensate" acceptor levels in the p-type material, creating an "intrinsic" region with high resistivity and correspondingly increasing the thickness of the sensitive layer (*c*).

able accomplishment. At the University of Rochester my colleagues and I are trying to employ the undepleted portion of the n layer in a thick n-i-p counter as the base for a thin gold-silicon surface-barrier counter. We hope in this way to dispense with the separate fragile wafer and at the same time convert the insensitive "window" layer of the n-i-p detector into a useful component of a combination counter.

Some nuclear radiations contain separate groups of particles so nearly equal in energy that even the best semiconductor counters cannot resolve them. In that case the groups must be sorted out spatially, by passing the beam through the field of a strong magnet. Each particle trajectory bends at a different angle, depending on the charge and momentum. The separated beams are normally detected by sending them

through special photographic emulsions, where they produce a visible track. But nuclear emulsions leave a great deal to be desired as particle detectors. Often the success or failure of a measurement hinges on some minor adjustment of the apparatus while the experiment is in progress. No such control is possible when working with nuclear emulsions because the experimenter cannot tell what is happening until the emulsions have been developed and scanned. Furthermore, scanning of the developed emulsions is in itself an onerous task.

A number of workers have tried replacing nuclear emulsions with banks of gas-filled or scintillation counters. To preserve the high resolution that the magnetic spectrograph provides, the individual counters must be quite narrow—on the order of two millimeters or less in thickness. Gas counters do not func-

tion well in such small dimensions. Scintillation counters can be cut to any size, but the need for individual photomultiplier tubes makes the array impossibly unwieldy. The semiconductor counter, on the other hand, being compact and self-contained, is eminently suited for assembly into arrays. Banks of 20 counters each have already been constructed and used successfully at the University of Rochester and at the University of Michigan.

The semiconductor counter has barely emerged from infancy. Many of its potentialities remain to be explored; some undoubtedly are yet to be thought of. As with so many technological aids to basic science, its development is a two-way street: while serving as a detector of nuclear radiation, it furnishes valuable information on the properties of semiconductors.

The Author

OLEXA-MYRON BILANIUK is assistant professor of physics at the University of Rochester. He was born in the western Ukraine, where he received his early schooling. After World War II Bilaniuk studied electrical engineering at the University of Louvain. Coming to the U.S. in 1951, he took degrees in physics, mathematics and nuclear physics at the University of Michigan, the last in 1957. Since 1958, when he joined the faculty of the University of Rochester, Bilaniuk has been engaged in research on the structure of atomic nuclei. The present article resulted from his work on the substitution of arrays of semiconductor counters for nuclear photographic emulsions. Last year Bilaniuk was a U.S. delegate at the United Nations conference on nuclear electronics in Belgrade. Currently he is doing research at the cyclotron facility in Buenos Aires at the invitation of the Argentine Atomic Energy Commission.

Bibliography

- HALBLEITER-SPERRSCHICHTZÄHLER.** W. Czulijs, H. D. Engler and H. Kuckuck in *Ergebnisse der Exakten Naturwissenschaften*, Vol. 34, pages 236–348, 1962.
- INTRODUCTION TO SEMICONDUCTOR PARTICLE DETECTORS.** W. L. Brown in *IRE Transactions on Nuclear Science*, Vol. NS-8, No. 1, pages 2–10; January, 1961.
- NUCLEAR EXPERIMENTATION WITH SEMICONDUCTOR DETECTORS.** D. A. Bromley in *IRE Transactions on Nuclear Science*, Vol. NS-9, No. 3, pages 135–154; June, 1962.
- SOLID-STATE DETECTORS FOR HIGH RESOLUTION NUCLEAR SPECTROSCOPY.** W. C. Parkinson and O. M. Bilaniuk in *The Review of Scientific Instruments*, Vol. 32, No. 10, pages 1136–1142; October, 1961.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound **SCIENTIFIC AMERICAN** Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

THE PLEIADES

by D. Nelson Limber

The stars in this familiar cluster move majestically about like bees in a swarm. These motions are a clue to how the stars came to be formed out of interstellar gas some 60 million years ago.

There are two ways of thinking about stars. On the one hand they are essentially point masses held together by gravity and hurtling through space like the molecules in an infinite container of gas. On the other they are nuclear furnaces in the process of evolutionary change. The astronomer's point of view depends on the problem he is considering. Sometimes it is useful to keep both aspects in mind. This is particularly true in the study of clusters of stars.

It seems likely that most stars were formed in clusters, condensing out of large local concentrations of gas and dust. In the ensuing millions or billions of years many of the groups have lost their cohesiveness and have been smeared out into the general stellar population. Any cluster that is still recognizable as such—and there are thousands in our galaxy—embodies, as it were, a double history: the spatial and dynamical evolution of the ensemble of its constituent stars considered as point masses and the internal and surface evolution of the individual member stars resulting from the nuclear processes taking place within them. As might be expected, each of these processes throws considerable light on the other, and the whole story turns out to be bigger than the sum of its parts.

For several reasons this twofold approach to the study of star clusters has not previously been attempted. My own work for the past few years at the Yerkes Observatory has been largely devoted to just such an attempt at understanding a cluster that is familiar to almost everyone: the Pleiades. To the naked eye the Pleiades consists of half a dozen loosely grouped stars in the constellation Taurus. The telescope reveals an association numbered at least in the hundreds, looking somewhat like a swarm of bees.

If the apparent motions of the stars were speeded up some billions of times, the analogy would be still closer, with the individual members darting this way and that while the swarm as a whole retained its coherence.

At one crucial point, however, the analogy breaks down. Knowing where all the individual bees in a swarm are now, and how fast they are going, would be of no help in predicting where they will be a minute hence or where they were a minute ago. But if the present mass, position and velocity of every star in the cluster were known (plus the distribution of any uncondensed interstellar gas and dust), an observer could, in principle, trace the detailed behavior of the group backward and forward in time as far as he wished—with certain restrictions. Unfortunately, although not surprisingly, no such complete specification of the present state of the cluster is available. From the fragmentary picture that we do have of the Pleiades cluster as it is now, we can hope to derive only an average dynamical history.

What are the details that can be made out? As far as stellar masses are concerned, these can be closely estimated from the observed luminosities and spectral properties of the stars identifiable as members of the cluster. Knowledge of the present positions and velocities of the individual member stars is less complete because neither their distances nor their velocities can be determined very accurately along the line of sight.

Incomplete as the data are, they do exhibit a striking pattern. There is a strong correlation between the distances of the member stars from the center of the cluster and their brightnesses, or, what is the same thing, their masses. The more massive stars are much more strongly concentrated toward the center than are the less massive ones. This may

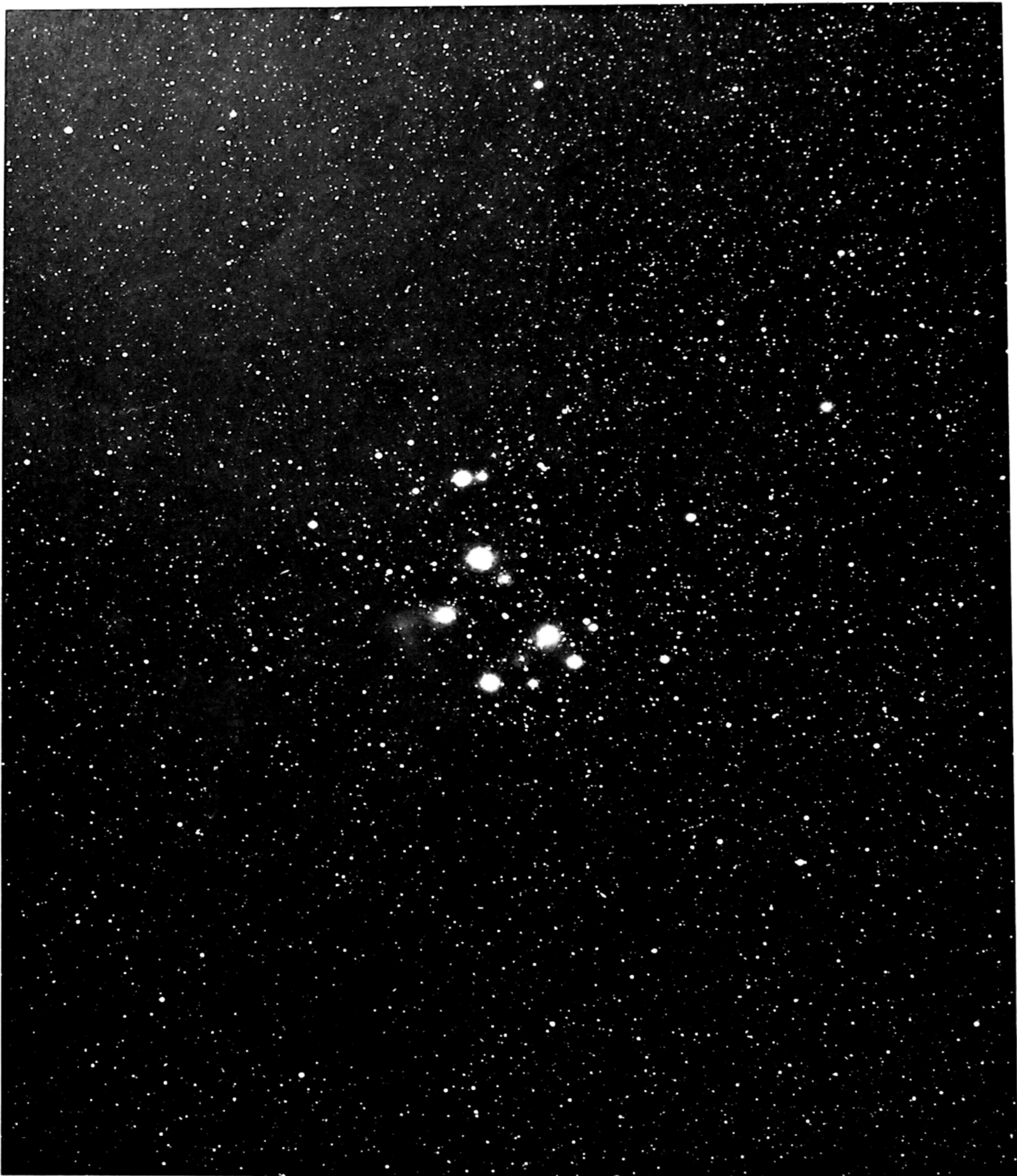
reflect an original tendency of heavier stars to form nearer the center. Or the pattern may have evolved through systematic dynamical effects after the stars had condensed. One of the major goals of my investigation has been to decide between these alternatives.

So far I have sketched the picture of the Pleiades that is actually visible in astronomical photographs. What about the part we cannot see? The cluster lies at a distance—about 400 light-years—such that stars much fainter than our sun are not visible at all, or at least cannot at present be identified as cluster members with any certainty. If there are many such stars, they will have a major effect on the dynamical behavior of the group as a whole. Moreover, there may also be a substantial mass of uncondensed gas, and it too would have to be taken into account. Some gas and dust actually shows up in photographs as glowing clouds around the brighter stars. But the clouds provide no reliable estimate of the total quantity of gas the cluster contains.

Fortunately radio astronomy now offers a direct method for estimating the

COLOR PHOTOGRAPH of the central region of the Pleiades on the opposite page was made with the 48-inch Schmidt telescope on Palomar Mountain. The bright star at top center and its fainter companion are Atlas and Pleione respectively. The central star is Alcyone; to the left and right below are Merope and Maia, both obscured by interstellar matter. The bright star at bottom center is Electra. The seventh member of the group, below and to the right of Maia, is Taygeta. The halos and spikes around the brighter stars result from the diffraction of light around the telescope's photographic-plate holder and its supports.



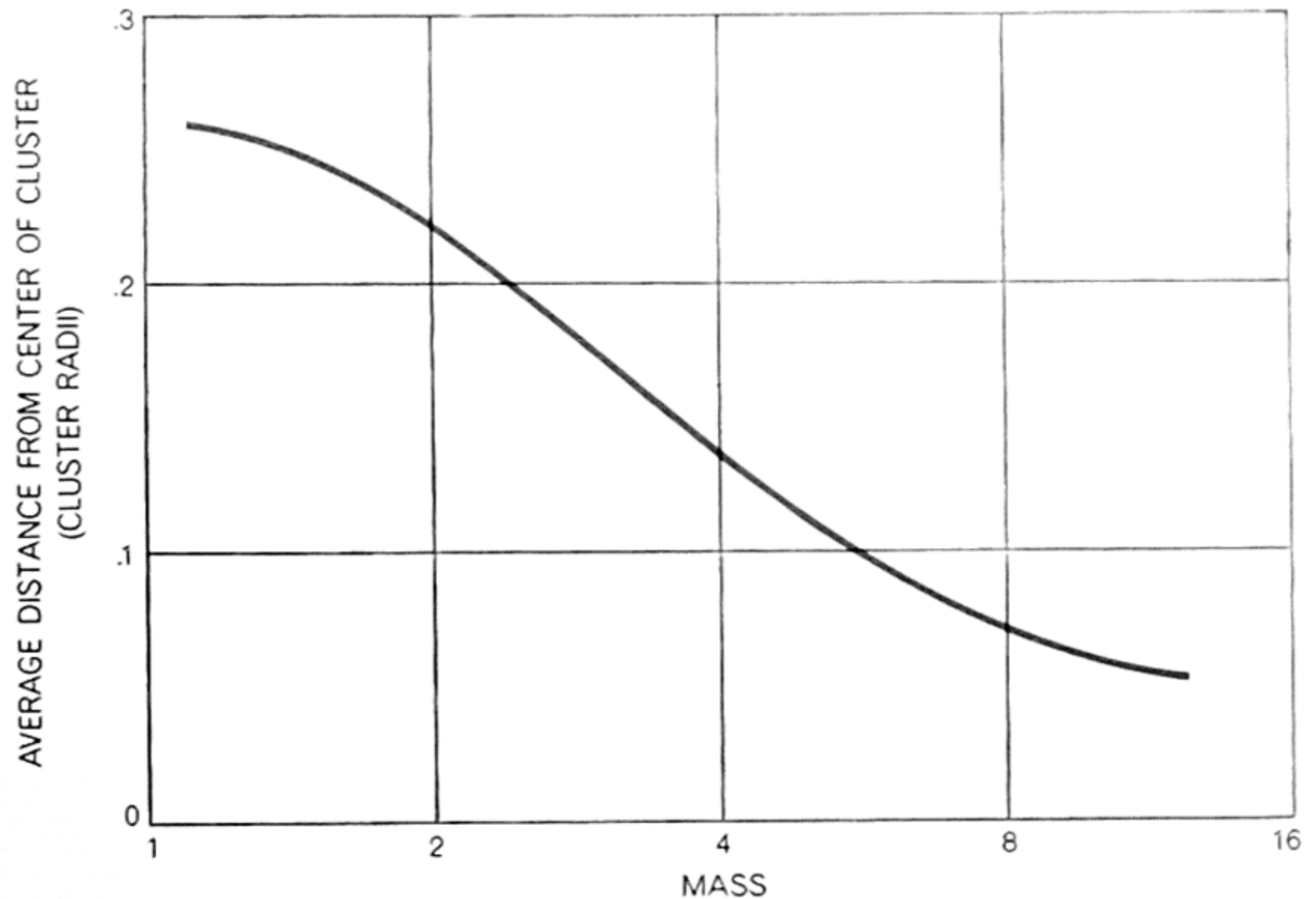


LARGER-FIELD PHOTOGRAPH, made with the 18-inch Schmidt telescope on Palomar Mountain, shows the configuration of the cluster's seven brightest stars and the faint, diffuse nebulosity

around Merope. The structure of the cluster itself is difficult to discern because many stars outside the cluster at varying distances in the foreground and background appear to fall within its limits.

amount of interstellar material if, as is probable, most of the material is in the form of neutral hydrogen gas. This method involves the measurement of the total energy emitted at a radio wavelength of 21 centimeters by neutral hydrogen [see "Radio Waves from Interstellar Hydrogen," by Harold I. Ewen; *SCIENTIFIC AMERICAN*, December, 1953]. Two groups of radio astronomers have recently measured the 21-centimeter radiation from the Pleiades. Their work offers a splendid example of the stumbling gait that so often characterizes scientific progress. One group reports a concentration of interstellar gas with a mass equal to 470 times the mass of the sun, which would be by no means negligible in the dynamical history of the cluster. The second group finds practically no gas at all! Needless to say, further studies will point to the source of error, wherever it may be, and eventually there will be a reliable figure for the mass of the interstellar matter. At that time it will also be possible to arrive at an unambiguous value for the total mass of the unseen stars. Meanwhile the latter figure can only be estimated on the alternative assumptions that one or the other group of radio astronomers is right. Presumably these two sets of calculations should at least roughly bracket the truth.

To deal mathematically with a large assemblage of stars, the astronomer finds it convenient to use the virial theorem, originally developed by physicists in a somewhat different form for dealing with the interacting molecules of a gas. In the form in which it is used here the theorem states that if an aggregation of stars, and possibly of interstellar matter, is neither explosively expanding nor rapidly collapsing, the disruptive tendency of the velocities of the individual members must be balanced on the average by the gravitational forces acting to collapse the cluster. The applicability of this theorem to the Pleiades in the present epoch seems well justified on several counts, both theoretical and empirical. Translated into mathematical terms, the virial theorem provides an equation relating the total stellar mass of the cluster, the total mass of the interstellar matter, the relative degrees of concentration toward the center of the cluster of the stars and of the interstellar matter, the dimensions of the cluster and an appropriate average velocity for the stars. Each of these quantities except for the total stellar mass can be estimated from observations. Therefore the equation



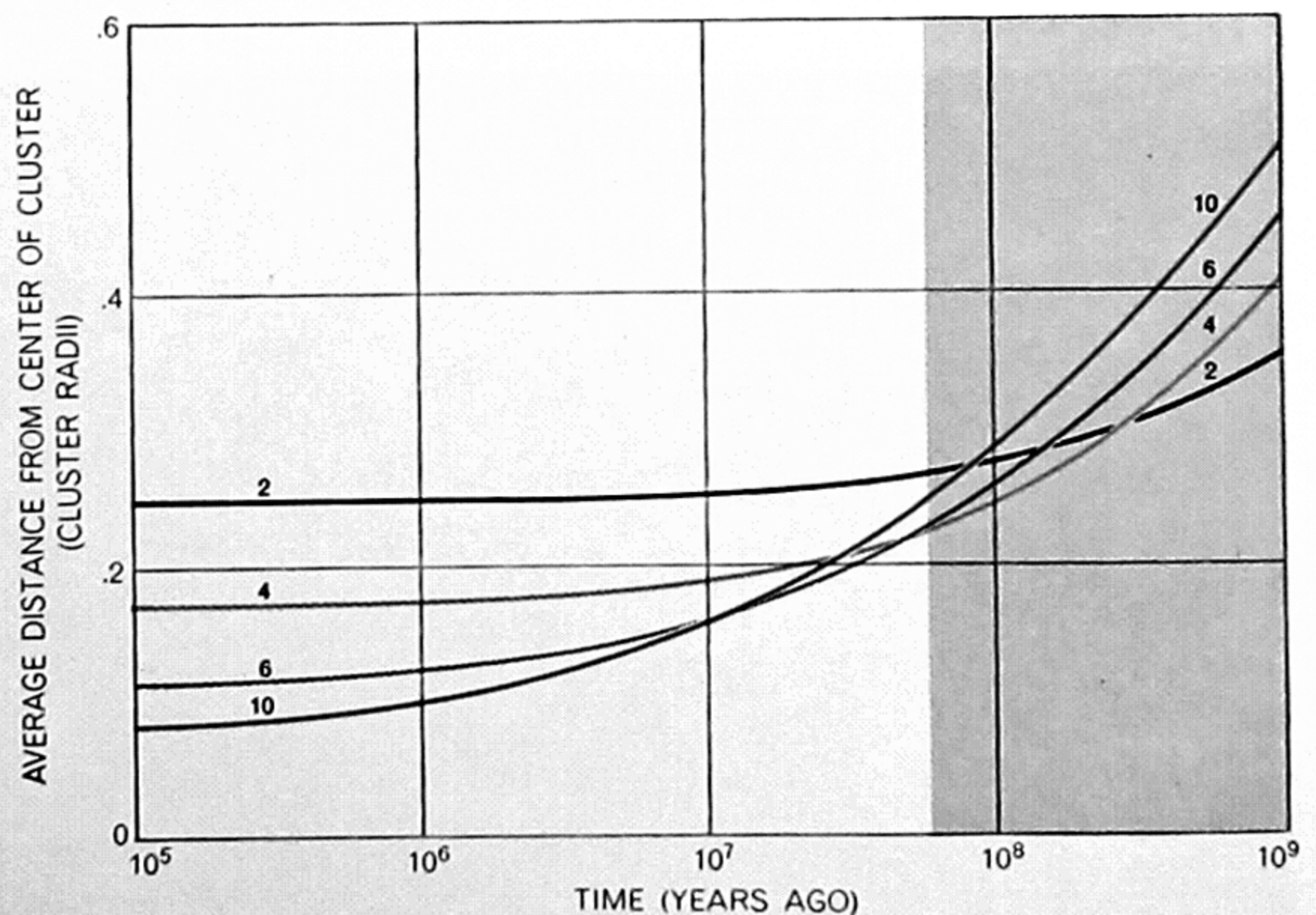
DISTRIBUTION OF STARS in the Pleiades cluster according to mass (in units of solar mass) reveals a preferential concentration of more massive stars toward the cluster center.

can be solved for its single unknown: the total weight of the member stars.

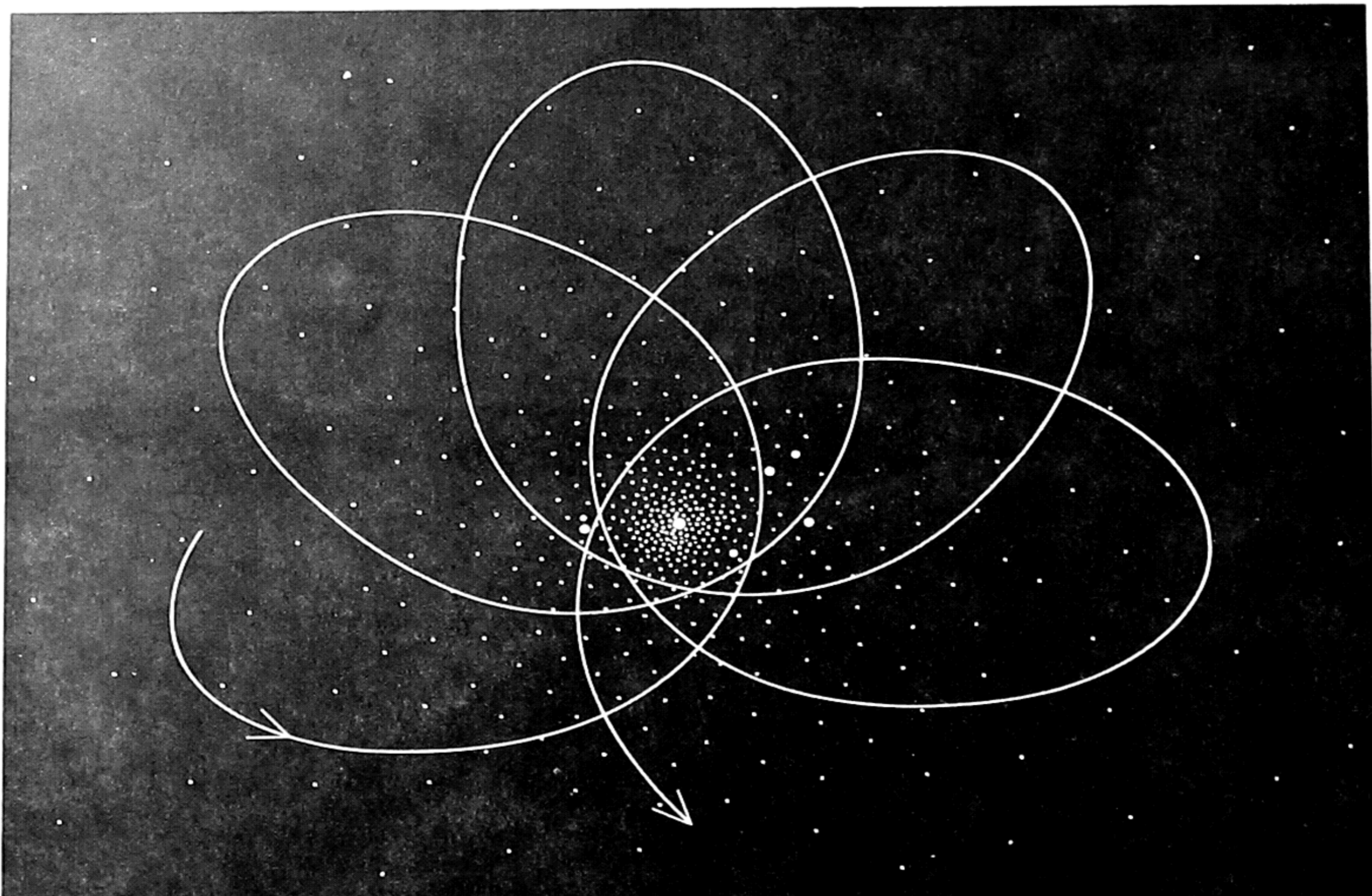
The answer turns out to be about 760 solar masses if there is a substantial amount of interstellar gas present within the cluster and 900 solar masses if there is not. The total mass of the visible members of the cluster is only about 350 solar masses. Therefore the Pleiades

must contain at least twice as much mass in the form of stars as is accounted for by the ones that have so far been identified as cluster members. Consideration of this unseen half of the cluster leads in a logical way to the other aspect of its history: the evolution of its individual stars.

As readers of this magazine are well

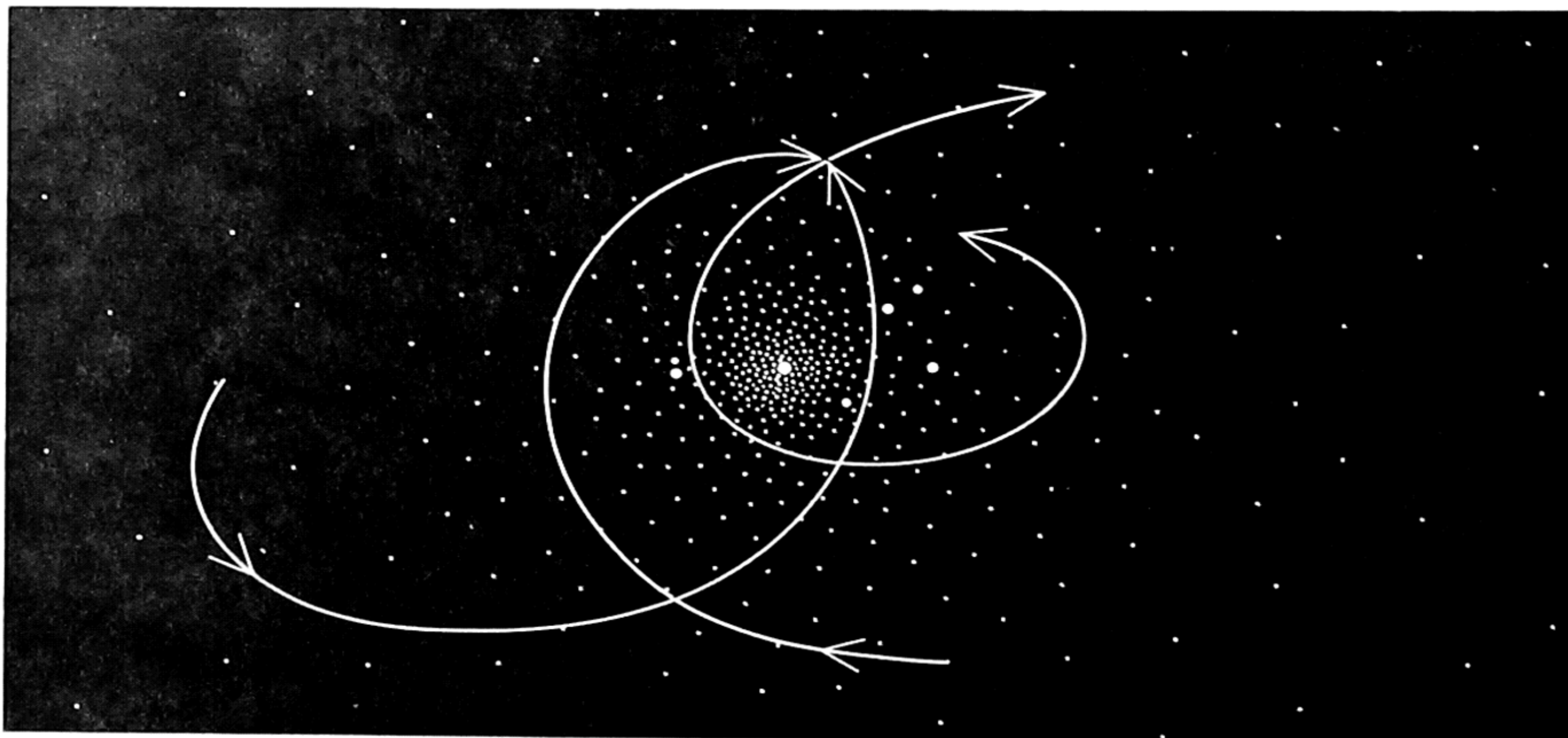


CHANGE IN DISTRIBUTION of stars during the past history of the cluster has brought the more massive stars closer to the center. The four curves shown here chart the movement of stars having two, four, six and ten times the mass of the sun. This distribution has not changed significantly in the past 100,000 years. The colored area represents the period of stellar history before the cluster formed, probably about 60 million years ago.



GRAVITATIONAL INTERACTION of stars in a cluster, in the limiting case where there are no gravitational encounters between

stars (see illustration below), will give a typical member star the type of motion around the center of the cluster that is shown here.



GRAVITATIONAL ENCOUNTER between two stars, which alters their orbits abruptly and drastically, is the second limiting case

of gravitational interaction of cluster stars. A more detailed representation of this encounter appears on the opposite page.

aware, the classic method for sorting out the evolutionary history of a group of stars is to plot the temperature of each star against its luminosity. When the visible members of the Pleiades are plotted in this way, with their surface temperatures measured along the horizontal axis (decreasing to the right) and their intrinsic brightness along the vertical axis (increasing upward), they all cluster around a line running from upper left to lower right known as the main sequence. Stars falling on the main sequence, as our sun does, are those that are deriving their energy from the conversion of hydrogen to helium in their innermost parts. The "initial main sequence" in the illustration on page 740 shows the theoretical positions of stars that have just finished contracting to this hydrogen-burning stage but have not yet consumed any appreciable portion of their hydrogen fuel.

When the temperature-luminosity diagram for the Pleiades is compared with plots of similar groupings of stars formed during the past several billion years, some interesting features emerge. The main-sequence stars in the Pleiades that are intrinsically brighter (and hence more massive) than our sun occur with about the same relative frequency as those in the other groups do. The brightest ones, it is true, tend to deviate from the initial main sequence, lying slightly above it. In the case of the fainter stars the picture is quite different. The Pleiades has far fewer of these on the main sequence than has come to be considered normal. The break between normal and abnormal distribution sets in for stars intrinsically only slightly fainter than our sun. Furthermore, just below this break point the Pleiades' stars begin to depart appreciably from the initial main sequence in the sense of falling above it. If one recalls the dynamical evidence for a large mass in the form of invisible stars, it is clear that many stars in the Pleiades must have not yet approached close enough to the normal main sequence to be observable at all.

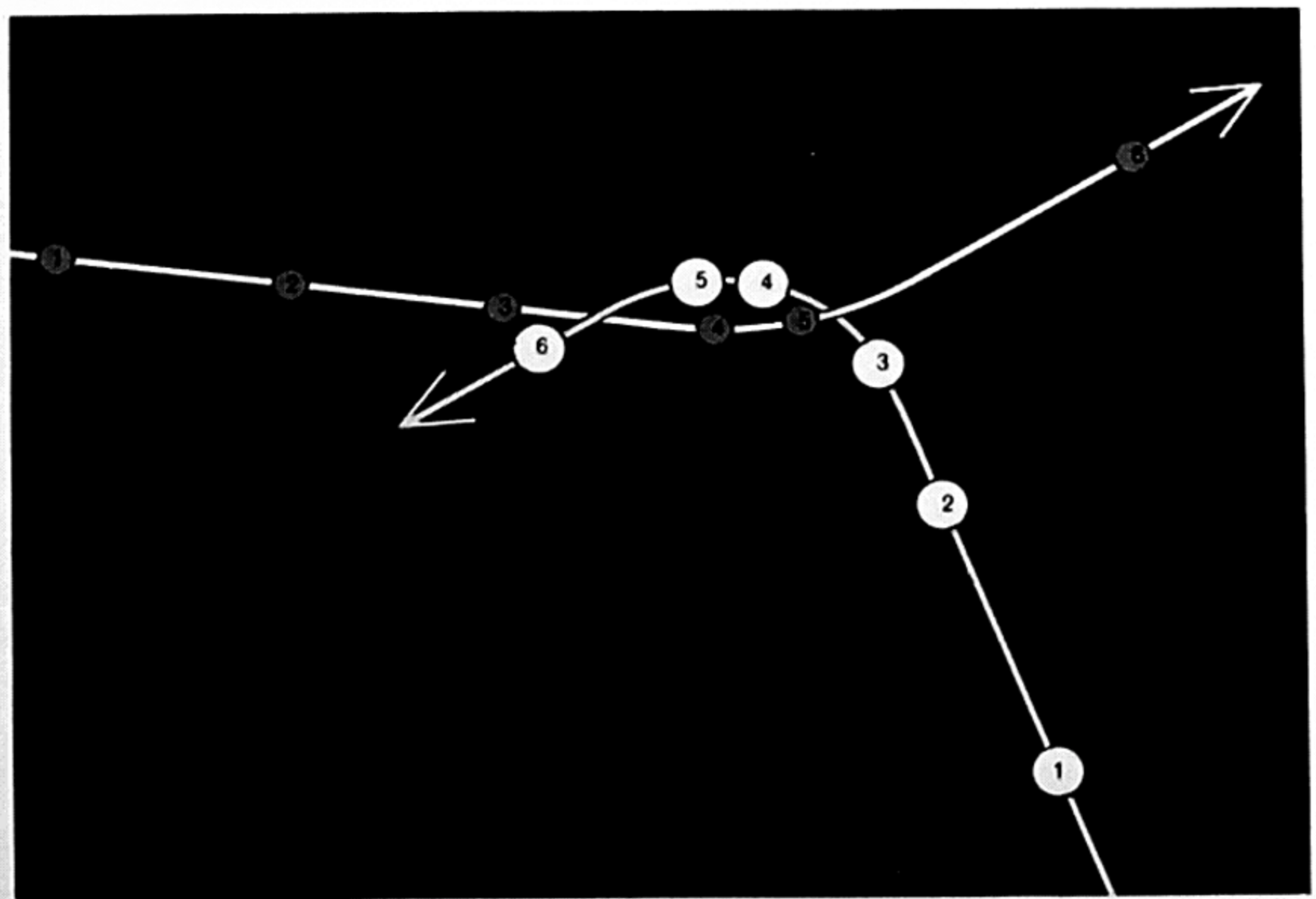
This interpretation is consistent with the present state of the theory of stellar evolution. According to the theory each star, once it has somehow begun to condense from the parent interstellar cloud, will continue to contract and to increase in temperature. Its starting point in the temperature-luminosity diagram will obviously be low on the luminosity axis and far to the right in the low-temperature region. (The temperature and

color of a star are connected, and a position far enough to the right on the temperature axis would correspond to a star that emits the bulk of its energy in the form of invisible infrared radiation. As the star heats up it first becomes red and then, if the contraction proceeds far enough, blue.) In the course of its contraction and heating it will move up and to the left. Eventually its internal temperature will increase enough to start the process of hydrogen-burning, and the star will attain a state of equilibrium, producing heat as fast as it is radiated away into space and for the time being undergoing no further gravitational collapse. At this point it will have arrived at its initial position on the main sequence [see illustration on page 740]. Just where it lands on the main sequence depends largely on its mass alone. Once it starts burning hydrogen the star continues the process until the supply of hydrogen in its hot central regions is exhausted. During the whole of this period it experiences only small changes in over-all structure and shifts its position on the temperature-luminosity diagram only slightly, moving somewhat upward and to the right from its position on the initial main sequence. Not until essentially all the hydrogen at its center is used up does the star undergo a major change that in a comparatively short time carries it far from the main sequence and into the region of yellow or red giants

in the upper right-hand section of the diagram.

The time that a star spends in contracting to the main sequence is only about 1 per cent of the time it spends in the hydrogen-burning stage on the main sequence. Both stages of evolution proceed more rapidly the more massive the star is. As the temperature-luminosity diagram of the Pleiades shows, the most massive stars have almost completed their stay near the main sequence and have just begun to move up and to the right. Theoretical calculations indicate that the whole process, from the time they first condensed from the interstellar material, must have taken some 60 million years.

Present indications are that all the stars in a cluster such as the Pleiades are about the same age, having begun their condensation and contraction toward the main sequence at approximately the same epoch. This suggests an explanation at one stroke of all the abnormalities in the temperature-luminosity diagram of the Pleiades: the deficiency of stars on the lower main sequence, the deviation of the fainter stars from the normal main sequence, and the large amount of stellar matter that cannot be seen at all. Suppose the time since the formation of the cluster is such that stars now at the break point between normal and abnormal sections have just had time to contract to the initial main sequence. This would mean that stars



SCHEMATIC DIAGRAM is an enlargement of the gravitational encounter shown in the bottom illustration on the opposite page. Such an encounter is not a collision, which is a highly improbable event. The two sets of numbers represent corresponding positions of the two stars at six distinct instants. One star (*color*) loses and the other gains kinetic energy.

of smaller mass, for the most part, have not had time to brighten enough even to approach the observationally accessible part of the diagram. The explanation is consistent with theory. Within the uncertainties of the theory, the contraction time for a star at the break point is indeed found to be about 60 million years—the same age that was derived independently from the study of the brightest stars. It should be pointed out that the detailed distribution of the stars below the break point in the diagram is not exactly what is predicted by the present theory for the contraction of stars to the main sequence. The reason for this discrepancy is not known. It may or may not be due simply to minor inadequacies in the current contraction theory.

So far a consideration of the present dynamical state of the Pleiades has led to the conclusion that it contains many unseen stars. An investigation of the stellar evolution of the cluster has shown why this should be so. It is now time to return to the dynamical history, armed with one all-important new fact: the cluster is some 60 million years old. In other words, it began to consist of separate, independently moving "point masses" 60 million years ago. It might be thought that any attempt to reconstruct the mechanical history of the past 60 million years would depend critically on the presence or absence of uncondensed interstellar gas. In this case, however, it turns out that the results are nearly the same on either of the two limiting assumptions mentioned earlier,

and so here the question is not crucial.

The dynamical history of the Pleiades is really the history of the gravitational interactions of its member stars. The net force experienced by a particular star, at a given instant, due to the totality of gravitational interactions can have two limiting forms. In one limit the star will on occasion find itself temporarily so close to one or more of the other stars that the force due to these chance near neighbors will greatly exceed the force resulting from all the other cluster members combined. If so, the net force on the star at that moment will be of a random sort, with no direct relation to the position of the star in the cluster as a whole. Such a star is said to be undergoing a gravitational encounter. (Actual collisions of stars in clusters are so



EARLY STAGE OF CLUSTER FORMATION in a gaseous nebula (NGC 6611) in the constellation Serpens was photographed with the 200-inch telescope on Palomar Mountain. The cluster, very prob-

ably still in the period of star formation, is at an earlier stage in its history than is the Pleiades at present. The nebulousity suggests the importance of interstellar gas to the early history of a cluster.

highly improbable as to play no role at all.) The other limit is one in which there are no stars close enough to the given star at the given time to exercise such a dominant role. In this case the net force experienced by the given star represents a smoothed-out average of the forces due to all the other members.

At any instant the force on a typical star in a cluster such as the Pleiades is almost always much closer to this second kind of limit than to the first. In this second limit there would be, strictly, no further changes of a systematic kind in the cluster after an initial adjustment period of several million years in which the member stars had time to move about through distances corresponding to the dimensions of the cluster. But any discrete gravitational encounters, few

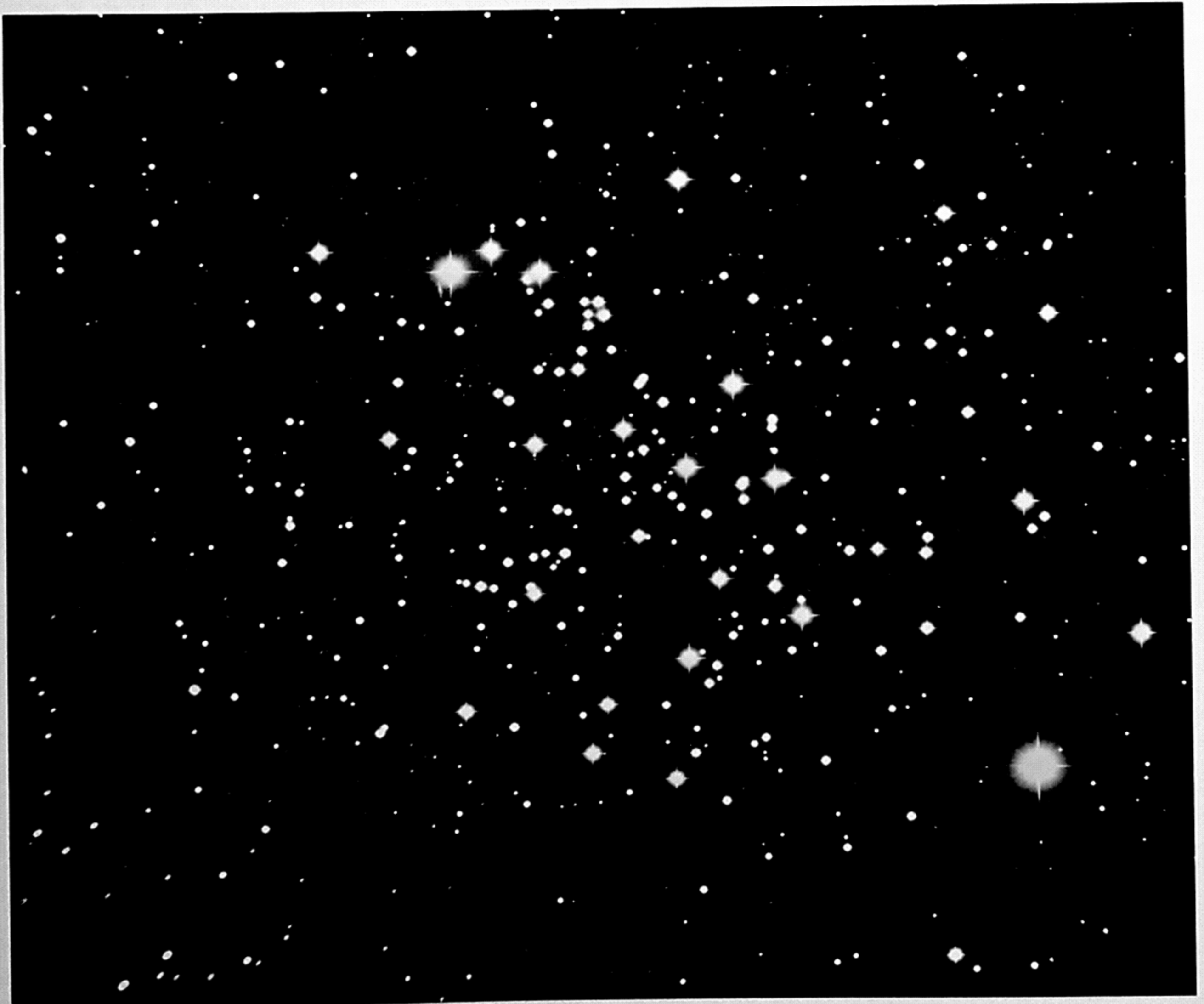
and weak as they may be, will alter this situation and give rise to systematic changes in the structure of the cluster as a whole.

The resulting systematic effects will be of two general types. Clearly there will be changes of a simple random, diffusive kind, which in the course of time will smooth out structural and dynamical details that may have existed in the initial distribution. Since it is impossible to trace back in time the motions of the cluster's individual stars, there can be no question of reconstructing the original detailed structure. But considering only the large-scale, over-all effects, it can be shown that there have not yet been important changes of this kind in the structure of the cluster.

In addition to diffusive effects on the

cluster as a whole, gravitational encounters can also be expected to produce another type of result that is systematically different for stars of different masses. This second effect comes about as a result of a quite general dynamical tendency on the part of interacting bodies to equalize their kinetic energies. The kinetic energy of a body is proportional to its mass and to the square of its velocity. Therefore if two stars of equal velocities but unequal masses undergo a gravitational encounter, the heavier star will, under average conditions, slow down and the lighter one will, under the same kind of average conditions, speed up in order to make their kinetic energies more nearly equal.

Now, the fact that the normal interaction of the stars in the Pleiades tends



STAR CLUSTER NGC 2682 in Cancer is, like the Pleiades, an "open" cluster. This photograph, made with the 200-inch telescope, provides a more representative view of such a cluster than does that

of the Pleiades on page 734, largely because of the smaller field in relation to the apparent size of the cluster and the consequent reduction in the number of stars that are not members of the group.

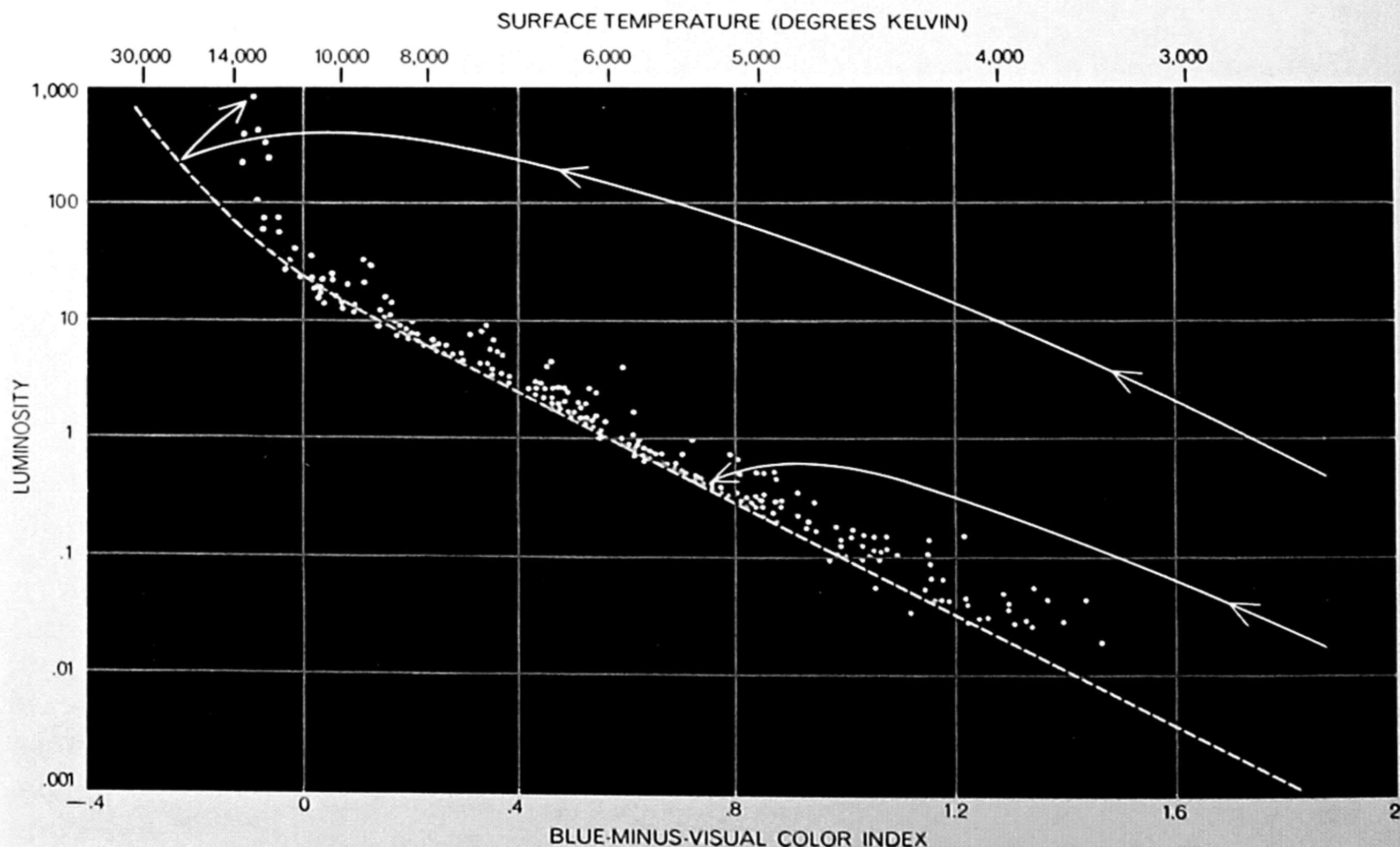
toward the second type of limit discussed above, in which each star is acted on by a smoothed-out average of the forces due to all the others, can be shown to mean that, at any given distance from the center of the cluster, stars will tend to have the same velocity regardless of their mass. Hence the average gravitational encounter, when it does occur, will in fact be between stars of nearly equal velocities. Consequently the heavier stars will tend to slow down and the lighter ones to speed up. This in turn means that the more massive stars will tend to fall in toward the center of the cluster until the speed gained in their fall becomes great enough to maintain them in equilibrium at a smaller average distance from the center, just as an artificial satellite, slowed a little by a retro rocket, say, falls a bit closer to the earth and speeds up. At the same time the lighter stars, having speeded up, will tend to move out to a greater average distance from the cluster's center, as a speeded-up satellite will swing outward into a larger orbit. (One can crudely envisage a large

weight as being at the center of mass of the cluster, and the individual members as, on the average, revolving around it under its gravitational attraction.) It should be emphasized that these changes do not take place in a discontinuous way but through a succession of smooth and gradual transitions that are always in process.

Clearly the mechanism just described will have tended to bring about the observed preferential concentration of more massive stars toward the center of the Pleiades. This suggests that we *begin* with the present concentration and work backward, applying a quantitative theory to determine at least approximately the distribution of member stars throughout the ages since the stars were formed. This calculation has been carried through, and the results are shown graphically in the bottom illustration on page 735. It is evident that 60 million years ago stars of different masses had very nearly the same average distance from the cluster's center. This result is important in the attempt to understand the processes that act in star formation,

since it indicates that stars of different masses do not form preferentially at different parts of the parent concentration of interstellar gas. This in turn suggests that the formation of stars of different masses does not result from large-scale differences in the physical conditions between different parts of the parent concentration. At best, though, the calculation gives only a clue and nothing like a complete answer to the intriguing problem of the factors involved in star formation.

This takes the story of the Pleiades as far back as it can now be told. But there is every reason to believe that a much more complete history will be written someday. Eventually one would hope to understand the process by which the original parent cloud condensed from the general gaseous substratum of our galaxy. And then the details of actual star formation must be filled in. In the course of extending the study, it will be surprising if some of the "facts" already uncovered do not turn out to need revision. There is, in any case, plenty of work to be done.



TEMPERATURE-LUMINOSITY DIAGRAM for the stars of the Pleiades cluster is discussed at length in the text. On the scale at left the luminosity of the sun equals unity. The color scale at bottom reads from blue at left to red at right. The broken colored line

marks the "initial main sequence." The shorter white line with arrows shows the development of an average star of relatively small mass up to the time of its arrival on this sequence; the longer line, the development of a more massive star up to and off the sequence.

The Author

D. NELSON LIMBER is associate professor of astronomy at the University of Chicago and a member of the staff of the university's Yerkes Observatory in Williams Bay, Wis. Limber's interest in astronomy began with his discovery, while walking home one night, of his "nearly total ignorance of the whole business." "I couldn't even find the Big Dipper," he continues, "and didn't have the slightest idea why the moon wasn't around or when I might expect it." Limber subsequently studied physics and astronomy at Ohio State University, where he received both his A.B. and M.Sc. in 1950. He acquired a Ph.D. in astronomy and astrophysics at the University of Chicago in 1953. Two years of postdoctoral work at the Princeton University Observatory were completed in 1957. Limber went to Chicago in 1958.

Bibliography

- THE COLOR-MAGNITUDE DIAGRAM OF THE PLEIADES CLUSTER: II. H. L. Johnson and R. I. Mitchell in *The Astrophysical Journal*, Vol. 128, No. 1, pages 31-40; July, 1958.
- THE DYNAMICS OF THE PLEIADES CLUSTER: I AND II. D. Nelson Limber in *The Astrophysical Journal*, Vol. 135, No. 1, pages 16-40 and 41-63; January, 1962.
- THE MILKY WAY. Bart J. Bok and Priscilla F. Bok. Harvard University Press, 1957.
- THE PHYSICAL MEMBERS OF THE "PLEIADES" GROUP. Robert Trumpler in *Lick Observatory Bulletin*, Vol. 10, No. 333, pages 110-119; 1921.
- PRINCIPLES OF STELLAR DYNAMICS. S. Chandrasekhar. Dover Publications, Inc., 1960. See Chapters II and IV.
- A 21-CM SURVEY AROUND THE PLEIADES. H. L. Helfer and H. E. Tatel in *The Astrophysical Journal*, Vol. 129, No. 3, pages 565-582; May, 1959.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

CHEMICAL TOPOLOGY

by Edel Wasserman

Although molecular rings are quite familiar, it has only recently been demonstrated that two such rings can be linked. It may even be possible to synthesize molecules in the form of knotted rings.

Organic molecules, having a great range of sizes, also have a rich variety of architectural forms. These days chemists are increasingly interested in the three-dimensional arrangement of the atoms of a molecule; this spatial structure is often an important element in determining the properties of a substance. In the past few years some of us at the Bell Telephone Laboratories have been examining a novel aspect of the organic structure problem, involving what might be called chemical topology.

What we have done, in brief, is to prepare and identify a structure we call a catenane (after *catena*, the Latin word for "chain"), which consists of two closed-ring molecules that are linked more mechanically than chemically, each simply threading the other. In one chemical sense a pair of closed rings constitutes the same structure, threaded or not. Each ring can be described independently of the other. Chemically they share no bonds and are in this sense independent. Topologically the cases differ, in the sense that two unjoined rings cannot be converted to a joined pair, or vice versa, without breaking one ring. A macroscopic pair of unjoined rings can lie flat on a table; a joined pair cannot.

What about the chemical properties of catenanes? That is what we are looking into. Certainly the breaking of a chemical bond in one of the rings is required to convert one form to the other. Therefore the two can be called topological isomers. (Isomers are molecules containing the same atoms in different configurations.) The chemical significance of topological isomers is a large, unexplored field that we and a few others are just beginning to investigate.

Most readers, in thinking of ring mole-

cules, probably will envisage the six-carbon benzene ring and others of about the same size that appear so profusely in structural formulas of organic compounds. These are not the kind of rings with which we are dealing. Their structure resembles that of a doughnut swollen to such an extent that the space in the middle has almost disappeared. With the addition of more carbon atoms to the straight molecule whose ends are to be joined to form a ring, some empty space appears. Not until the hole is big enough to admit a second straight segment is it possible to form a catenane.

Working with accurately scaled models of CH_2 groups, the chief units in the precursor straight molecules, we were able to determine that a ring would have to contain 20 CH_2 units before another unit could thread it without having to overcome large repulsive forces from the ring molecules, as would be the case in a tight fit [see illustration on page 744]. Anything bigger, of course, would make the threading easier, but larger molecules are harder to make and harder to handle because of their decreased solubility. Our compromise was a 34-carbon ring, which makes what we designated a 34, 34 catenane. It should be mentioned that our rings, and most large rings, have no double bonds between carbon atoms as the benzene ring does.

We were by no means the first to think of making linked-ring compounds. More than 50 years ago the German chemist Richard Willstätter considered the possibility. The ring molecules known at the time, however, were too small. In the 1920's the Swiss chemist Leopold Ruzicka developed methods for making rings with as many as 34 carbon atoms, but these methods

yielded only a tiny proportion of large rings. Finally in 1947 two other Swiss chemists, V. Prelog and M. Stoll, independently developed a general method of making large rings in quantity. Any of the several methods available today for producing large rings must also yield some catenanes, and by the same process as ours. We do not push straight molecules through rings; we simply rely on the law of averages to do it for us a small percentage of the time when rings and straight molecules are mixed. According to calculations by H. L. Frisch of the Bell Laboratories a few per cent of the 34-carbon rings can be expected to interlock.

What we feel we have done is to increase the yield of catenanes to a detectable amount and, for the first time, to prove that such interlocked rings are present. Our method for forming rings is the same as that of Prelog and Stoll. In principle we start with a string of carbon atoms that has an acid group (COOH) at each end. This string is combined with methyl alcohol to convert each end group to form a COOCH_3 , making what is known as a diester [see illustration on page 745]. Then in an atmosphere of inert gas the diester is slowly added to liquid xylene (a non-reactive solvent) that contains finely divided sodium. Now acetic acid is added and reacts to produce a circular molecule called an acyloin. This molecule has a hydroxyl group (OH) and an oxygen atom (O) on the two carbon atoms that were the ends of the linear molecule. The rest of the molecule consists of carbon and hydrogen. There are several ways to prove that the molecules have in fact formed rings, such as cleaving them once and showing that the number of molecules remains the same.

Every solution of acyloin rings un-

doubtedly contains some that have accidentally linked, but they would be very difficult to detect. What one needs is a pair of slightly different rings that can be identified when they become linked. To this end we prepare a batch of acyloin rings and then alter all of them, keeping the rings intact. We treat them with hydrochloric acid in which the hydrogen has been replaced by heavy hydrogen, or deuterium. The result is both the removal of the oxygen atoms from the acyloin ring and the substitution of deu-

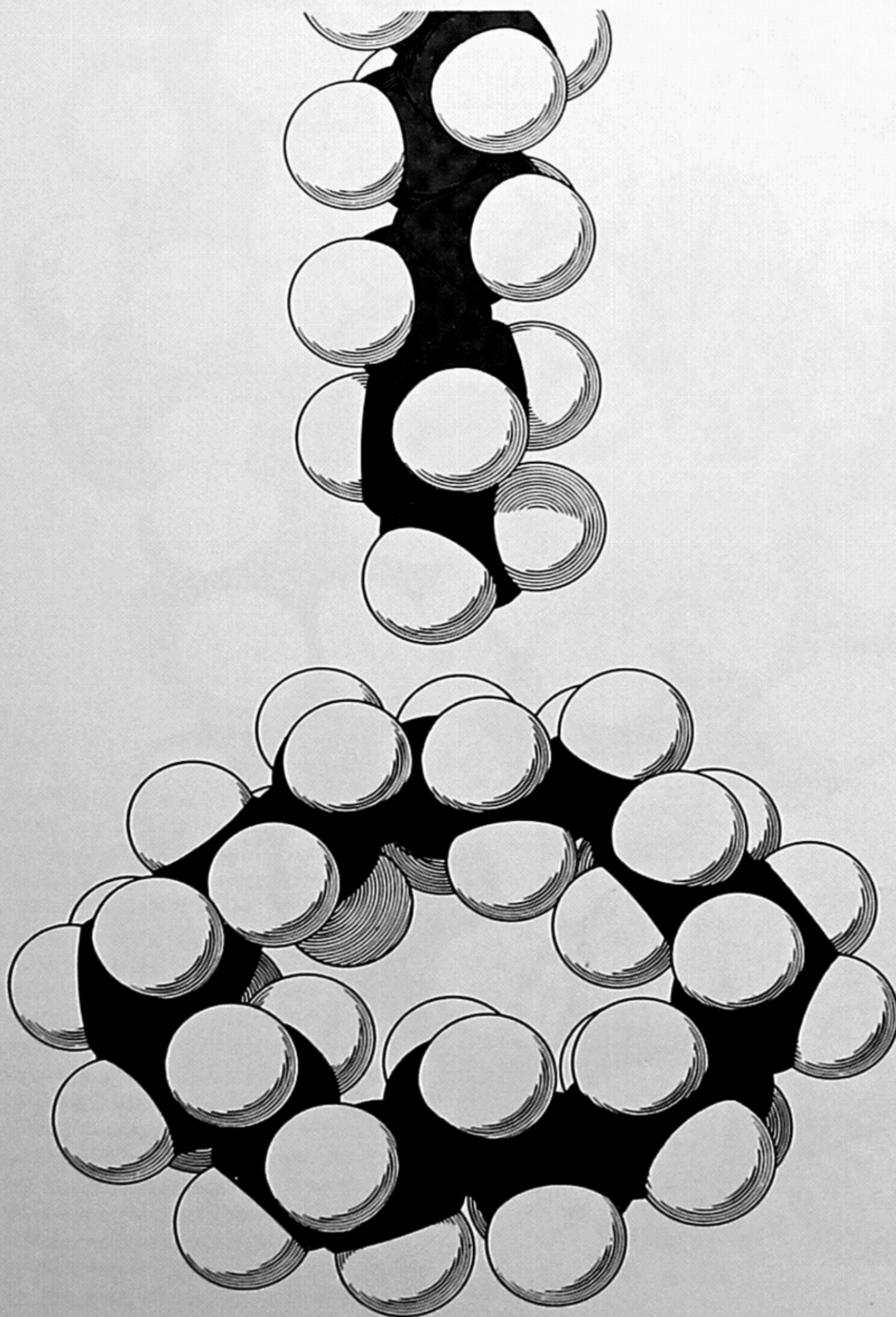
terium for approximately five of the 68 hydrogen atoms in the ring. We call the altered rings A rings; the acyloin rings we call B. The removal of oxygen turns the ring into a comparatively inert compound known as a cycloparaffin. The deuterium content of the A ring means that it responds somewhat differently to infrared radiation than does an ordinary all-hydrogen cycloparaffin ring. Because the deuterium atom is heavier than the hydrogen atom, it vibrates more slowly in response to infrared rays, which means

that the carbon-deuterium bonds absorb infrared radiation of lower frequency (longer wavelength) than do carbon-hydrogen bonds. The difference in activity is revealed clearly by a spectrograph [see illustrations on page 747].

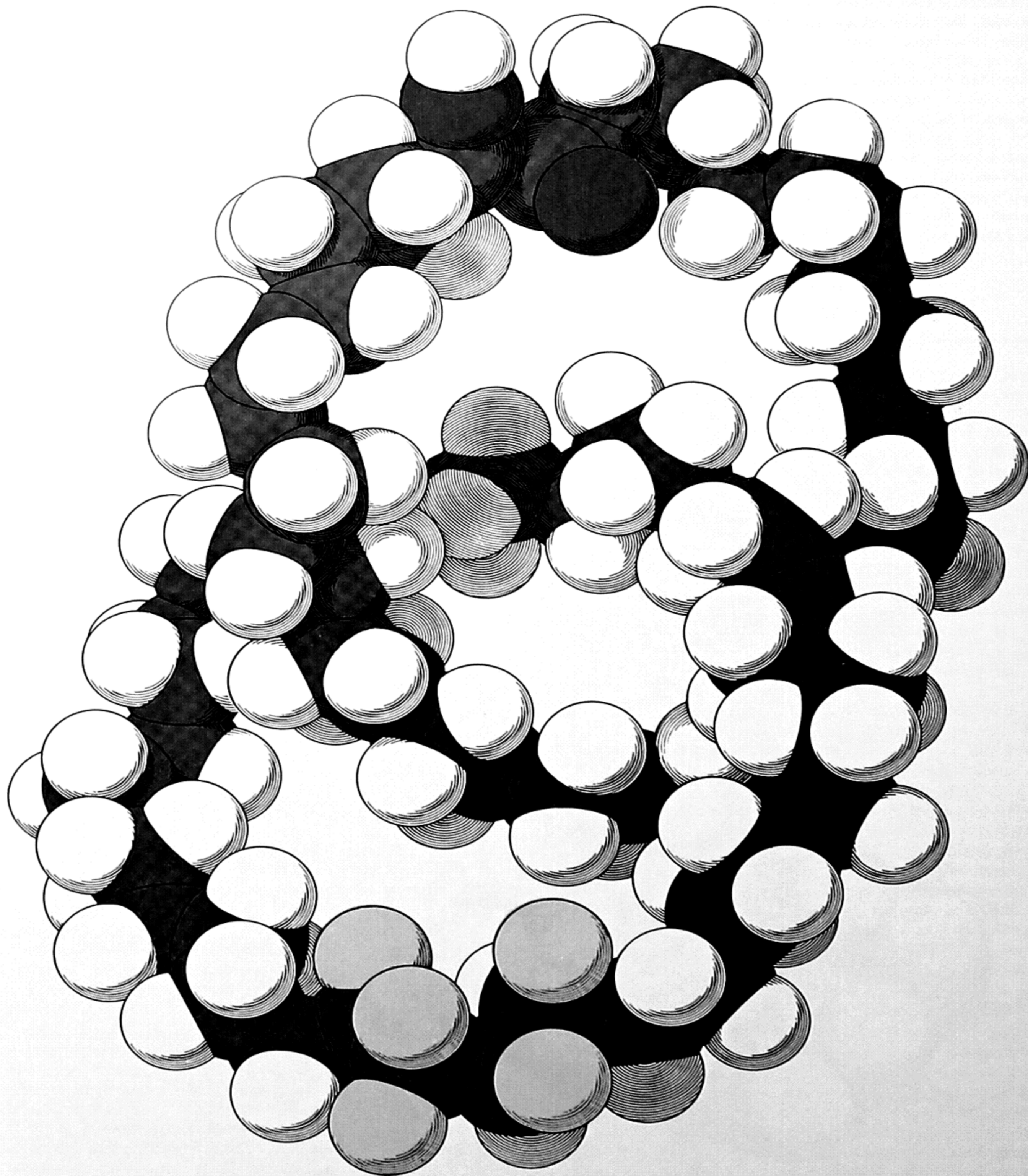
Having formed the deuterated A rings, we use them to replace much of the xylene as the solvent, add more of our linear C_{34} diester to this solution and convert the diester to B rings. Some of the linear molecules thread A rings before we change them into acyloin rings. These B rings contain oxygen atoms, of which A has been deprived. The final product of the reaction consists of the deuterated ring A, ring B and the A, B catenane.

We separate the A, B catenane from the other molecules by the chromatographic process of passing the solution containing the various rings through a glass tube filled with a powdered solid, such as alumina or silica gel. Ring A has only carbon and hydrogen atoms (including the deuterium), and these are not attracted to the solid. The ring washes out of the tube when we pour pentane, an inert solvent, through. The oxygen atoms on ring B, however, attach themselves to the particles of powder, and B (including A, B catenane) remains in the column. Now we pour into the column an active solvent, such as ether or methyl alcohol, that contains oxygen. This displaces B (and A, B catenane) and carries it out of the column in solution [see illustrations on page 746]. The infrared spectrum of this second solution indicates the presence of deuterium. The amount suggests that of the 10 grams of A in the original reaction mixture, about one milligram (.01 per cent) is associated with the 100 milligrams of ring B that were formed.

The presence of deuterium along with B rings is consistent with the presence of an A, B catenane. Other possibilities, however, must be excluded. For example, some A may have been oxidized by air (although this is a very slow process), during the chromatographic separation. Addition of oxygen in this way would make A stick to the powder. To test this possibility we unlock the B rings by chemical reactions that yield the linear diacid from which ring B originally came. The reactions have no effect on ring A; it is simply freed from the catenane. Chromatography of the solution produces free ring A that, further testing shows, contains no oxygen atoms. Its melting point and infrared spectrum,

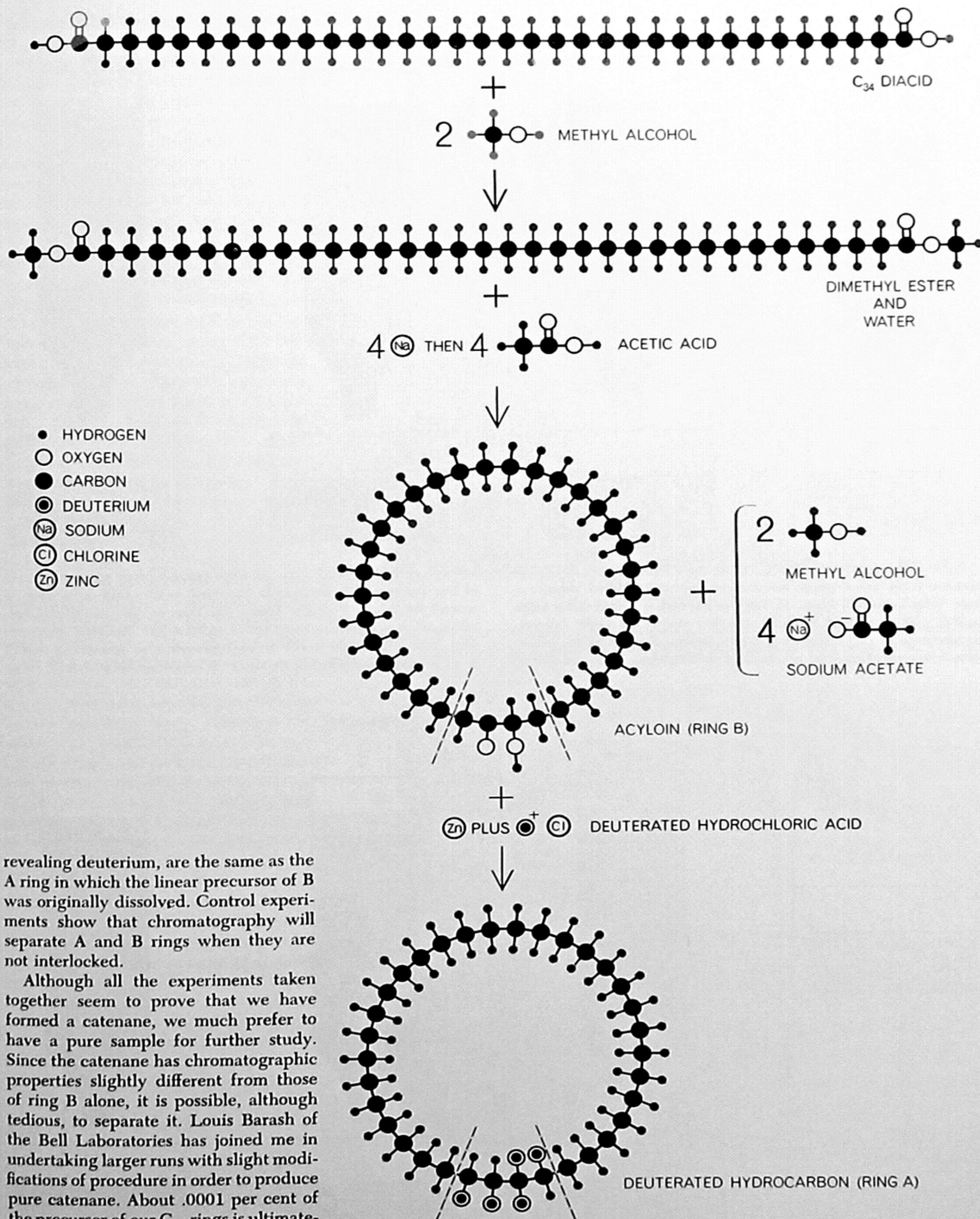


SMALLEST RING that can be threaded by another hydrocarbon molecule, when hole is wide open, contains 20 carbon atoms. As part of his investigations, the author builds such models, in which the "atoms" have the same relative sizes as in the chemical compounds.



TWO LINKED RING MOLECULES are topological isomers of an unlinked but otherwise identical molecular pair. The two shown here each contain 34 carbon atoms (*gray*). Ring A (*the lower*) is a paraffin, a relatively inert compound consisting of carbon and hydrogen (*white*) including five atoms of heavy hydrogen, or deu-

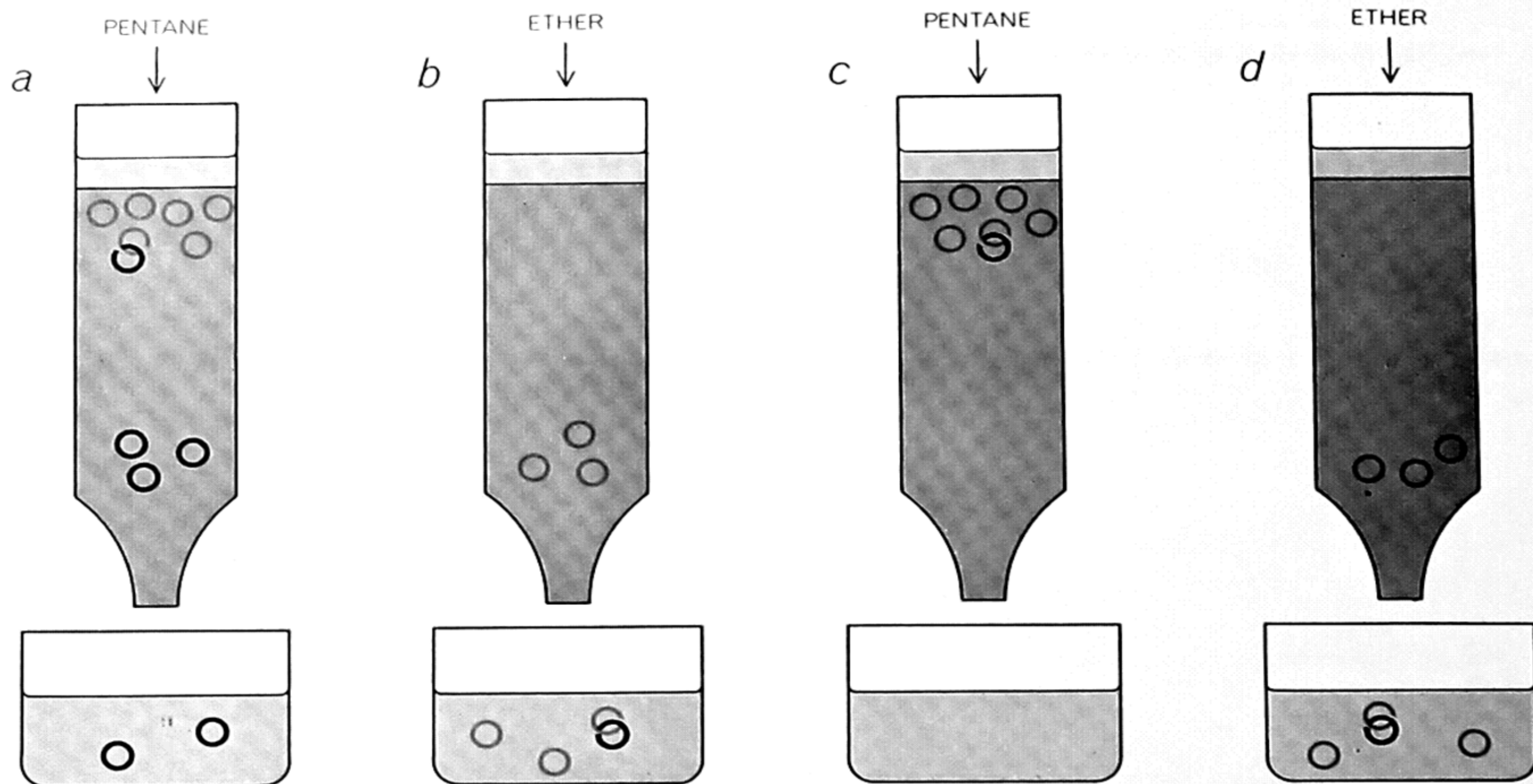
terium (*color*). Ring B (*the upper*) is an acyloin, an active compound that has two oxygen atoms (*color*). This 34, 34 catenane is the compound that the author has constructed, identified and isolated. The rings are not rigid. They change shape constantly, and most of the time the holes are not so large as those shown here.



revealing deuterium, are the same as the A ring in which the linear precursor of B was originally dissolved. Control experiments show that chromatography will separate A and B rings when they are not interlocked.

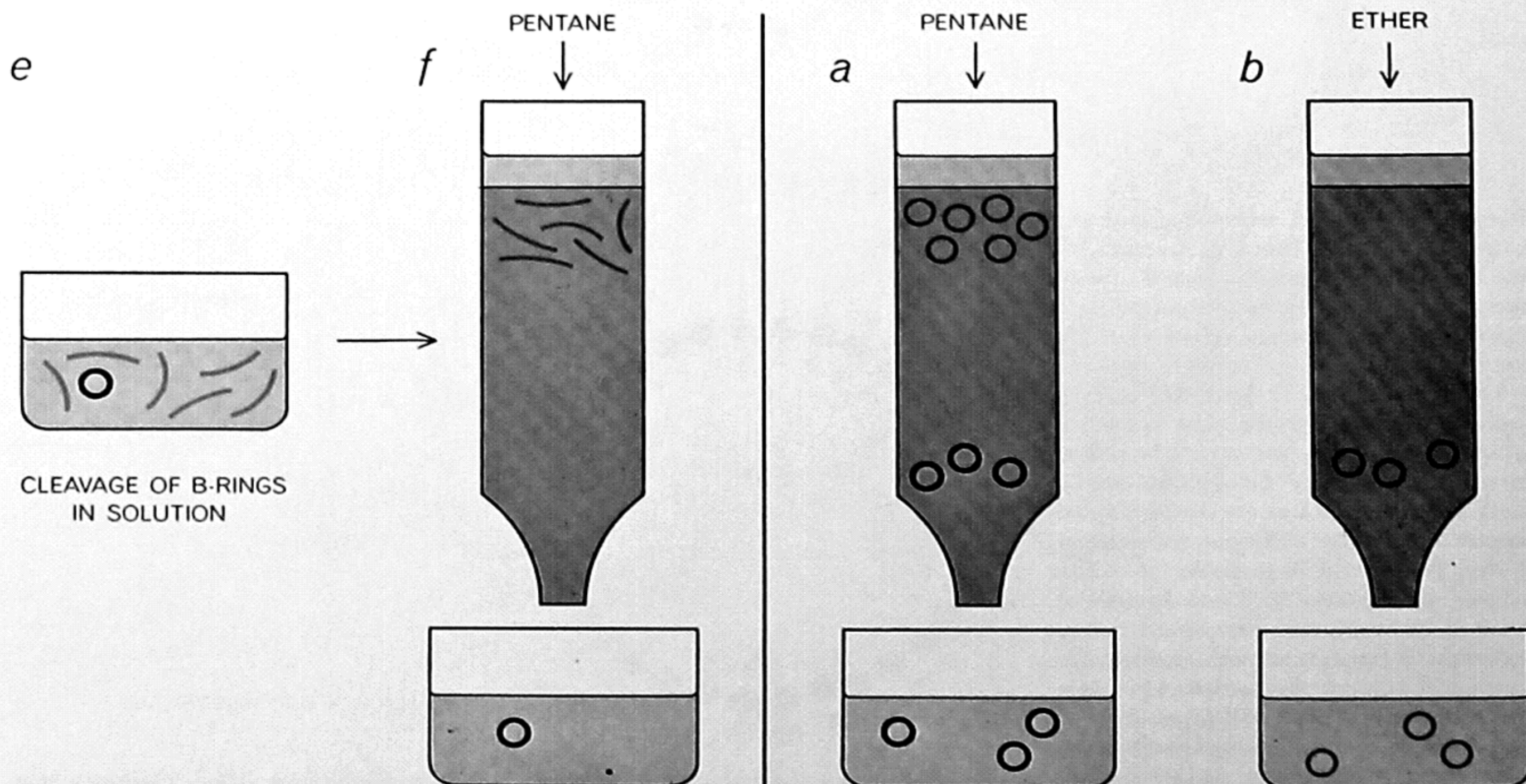
Although all the experiments taken together seem to prove that we have formed a catenane, we much prefer to have a pure sample for further study. Since the catenane has chromatographic properties slightly different from those of ring B alone, it is possible, although tedious, to separate it. Louis Barash of the Bell Laboratories has joined me in undertaking larger runs with slight modifications of procedure in order to produce pure catenane. About .0001 per cent of the precursor of our C₃₄ rings is ultimately converted to catenane. At this writing we have a few milligrams (totaling perhaps a 5,000th of an ounce) of isolated catenane. About all we can say of it is that it seems to be an oil. In the near

CHEMICAL REACTIONS that produce ring molecules begin with a linear molecule that has an acid group at each end (*top*). The broken lines in the two rings mark small reactive regions that were previously the ends of the linear molecules. Other chemical reactions can turn the upper ring back into linear diacid. Schematic diagrams on this page do not show true appearance of the molecules. Holes in the real ring molecules are much smaller.



CHROMATOGRAPHIC PROCESS is used to separate the A, B catenane from free A rings. Solution of catenane (*linked rings*), A rings (*black*) and B rings (*color*) is poured (*a*) into glass tube packed with powdered alumina (*darker gray*). An inert solvent, pentane, washes out A rings but leaves catenane and B rings, which contain oxygen atoms and are held by the alumina. Ether, an active

solvent, displaces the B rings and the catenane and carries them out of the column (*b*). Process is repeated, using only molecules removed by ether. Pentane comes out free of ring molecules (*c*), while ether again takes rings out of column (*d*). Infrared spectrometry (not illustrated here) proves presence of deuterium, which is in A ring only. Obviously some A rings are locked to B rings.



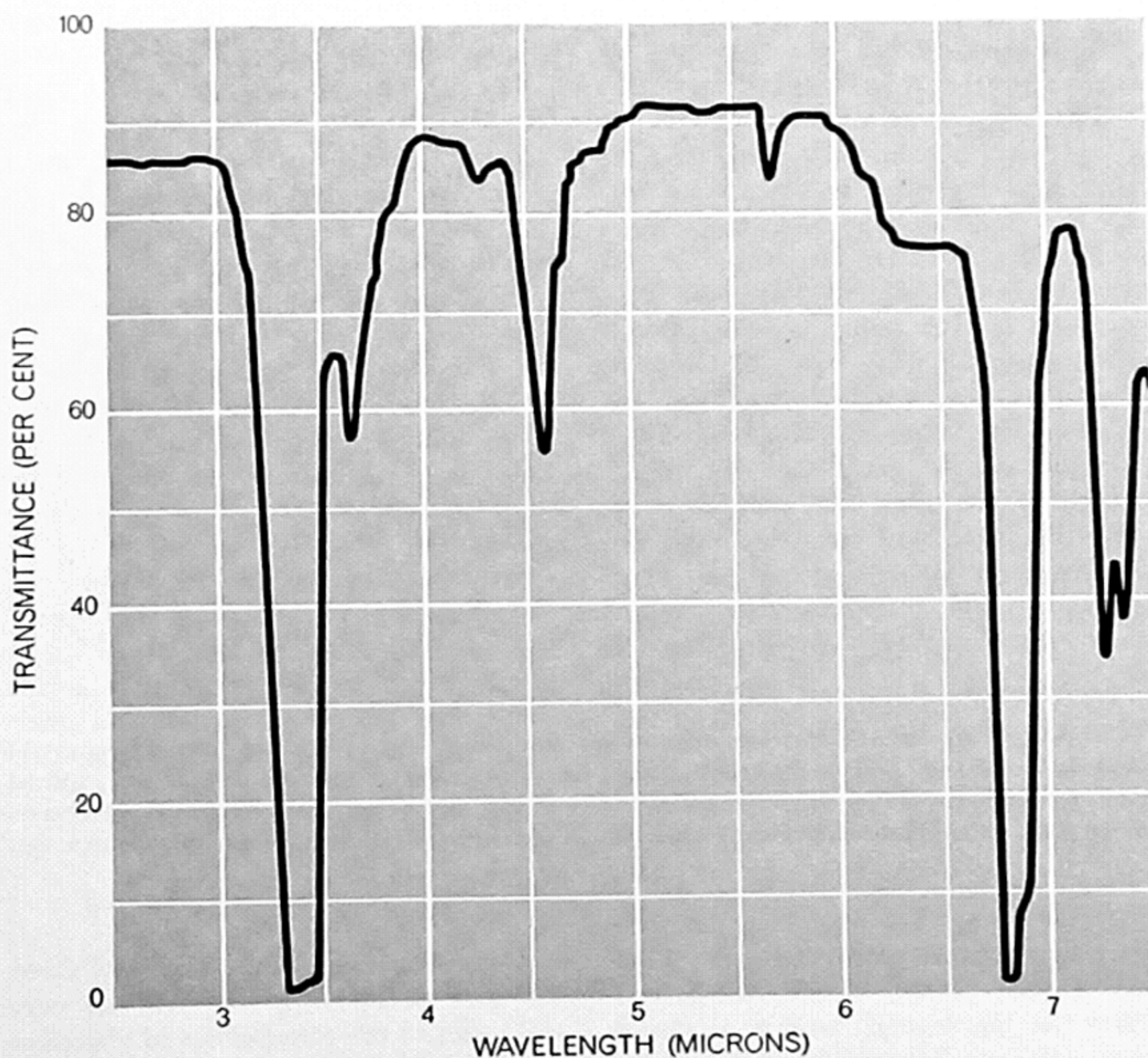
TEST CONTINUES with unlocking of B rings (*e*), which frees A ring from catenane. Pentane then washes the molecule containing deuterium (A ring) out of the chromatographic column, leaving linear diacid that resulted from the cleavage of the B rings (*f*).

CONTROL TEST shows that inert solvent, pentane, removes A rings from mixture of A and B rings (*a*). No catenane is present. Then the active solvent, ether, removes the B rings from the column (*b*). This experiment works under a variety of conditions.

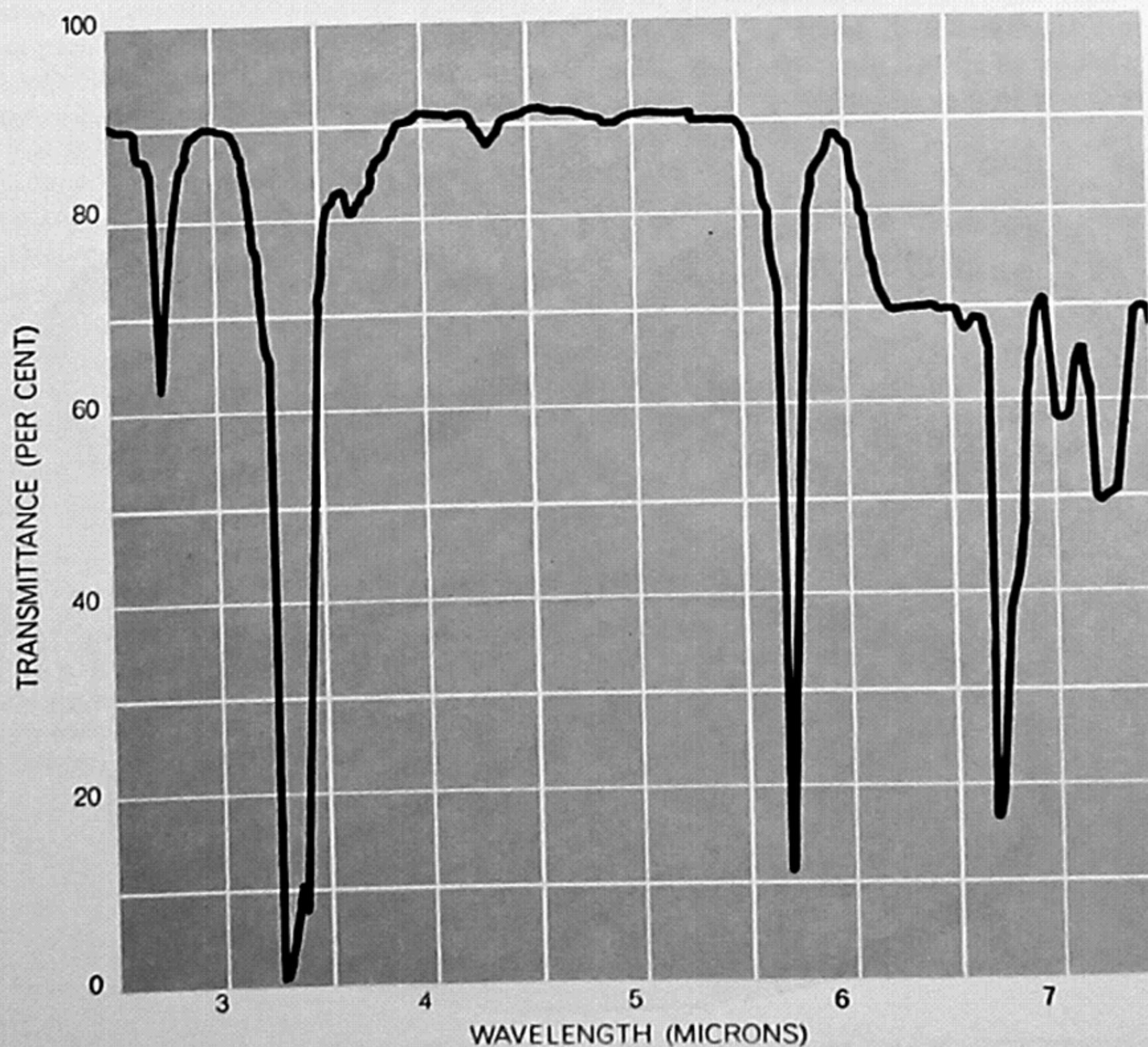
future we should be able to learn something about its properties. (It may be possible to produce much higher yields by other methods. David M. Lemal of the University of Wisconsin and Lester Friedman of the Case Institute of Technology are trying to synthesize a molecule in which both rings are attached to a central core in such a way that when the core is removed by chemical reactions, a pair of interlocked rings remains.) When we have enough catenane, we will be able to perform some interesting tests. One thing we will do is to convert the B ring in the combination into another inert A ring, so that our two links will be identical except for the deuterium in one, which should have no appreciable effect. Then we can compare the properties of an A, A catenane with a solution of unlinked A rings. The differences, if any, that we find will be solely due to the linking in the catenane. It will be much easier to make a valid comparison when all the rings are of one type than it is when we compare the more complex A, B catenane with a mixture of unlinked A and B rings.

In addition to the two interlocked rings, other forms of catenane are possible. One would arise if the precursor of ring B threads A twice before closing. Molecular scale models indicate that the minimum size for this structure would be a C_{33} molecule, only one less than the C_{34} we have been using. Therefore the amount of double-threaded form we actually get should be minute compared with that of the simple chain. If the linear diester threads two rings before closure, a three-link chain will form. Since only about 1 per cent of B joins in a catenane, less than .01 per cent should thread two A rings. Borromean rings, in which all three rings are joined but no two of them [see top illustration on page 748], require a minimum string of 30 carbons and involve a considerable restriction on the motion of the loops—both unfavorable for the construction of such a complex form. Other possibilities are even more unlikely. We believe that our linked rings are more than 99 per cent simple two-link catenane.

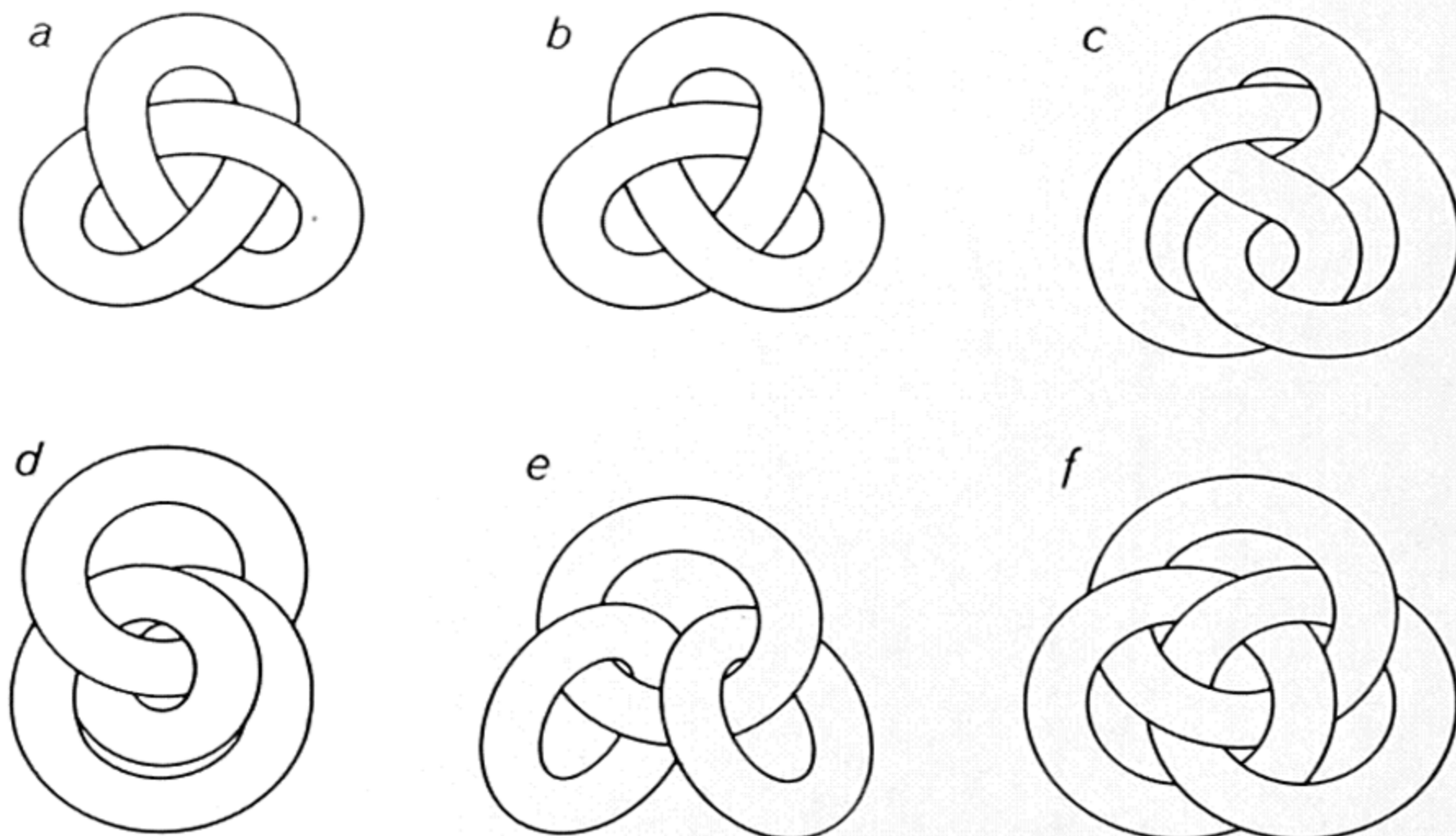
One interesting question about the catenane is whether it is one molecule or two. In ordinary "geometrical" chemistry a molecule consists of atoms interconnected by chemical bonds. This is not the case with our material. We must nonetheless break a chemical bond to split the catenane apart. Perhaps it is best to say that a "topological bond" exists between the rings. The bond is not the property of any pair of atoms but of



INFRARED SPECTRUM shows amount of infrared radiation transmitted at different wavelengths by solution of ring A. Deuterium in the ring absorbs at wavelength near 4.6 microns.



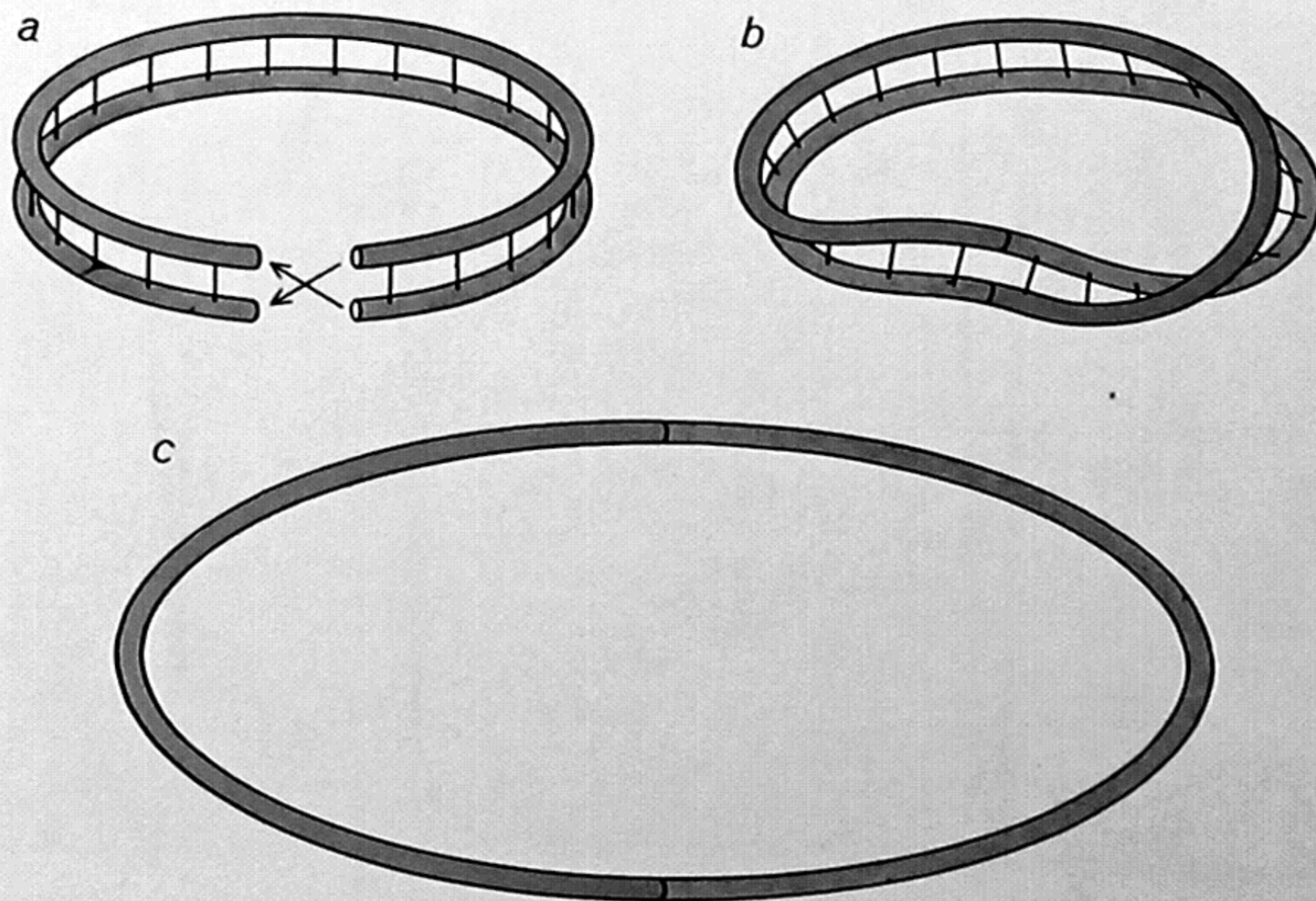
ANOTHER SPECTRUM is obtained from solution containing only ring B, which has no deuterium. This compound absorbs not at 4.6 microns but rather at 2.7 and 5.8 microns.



TOPOLOGICAL ISOMERS when ends of linear molecules are joined to form a ring could include "left-hand" and "right-hand" trefoil knots (*a* and *b*), figure eight knot (*c*), linked rings with one threading the other twice (*d*), one ring threading two others (*e*) and Borromean rings (a well-known brewer's symbol), in which three rings are joined but no two (*f*).

the two complete rings. It breaks when any carbon-carbon bond is cleaved. Since the topological bond is as strong as an ordinary chemical bond, we prefer to regard our catenane as a single molecule. Determination of its molecular weight and other properties, such as its rate of diffusion in solution, will show whether or not we are right in thinking of it as a single molecule.

Catenanes and combinations of three or more linked rings are not the only examples of the possibilities of chemical topology. H. L. Frisch of the Bell Laboratories and Norman Martin Van Gulick of E. I. du Pont de Nemours and Company have independently pointed out that a single large molecule, ring or otherwise, can have a knot in it. Only a ring, however, would have to be cut



DOUBLE-STRANDED MOLECULES with cross-links (like rungs of a ladder) might be joined as in *a* to produce a molecular Möbius strip (*b*) with half-twist about the long axis. Cleavage of all the cross-links would produce the large single-ring molecule shown in *c*.

to be untied, and so only a ring is topologically different from the unknotted form. Scale models show the size of ring needed for various knots. Rings with 50 or more carbon atoms can have a simple overhand knot, or trefoil (so named because it has three cross-over points). No other knot is possible until the molecule has 74 carbons, when a figure eight, the only four-fold knot, may appear. There are four different six-fold knots, 21 eight-fold knots and 133 10-fold knots. To figure out the minimum number of carbons required for each knot requires only patience and enough atomic models.

Although the probability of forming knots increases with ring size, separation of the different kinds of knots in a mixture and their identification is a nightmare we prefer to avoid. In order to have only one alternative to the simple ring, we have prepared a C_{66} linear diester for a knot test. We guess that statistically about .1 per cent of these "string" molecules may have overhand knots in them after conversion to rings. Separation and identification is going to be difficult. We have no analogue of the deuterium label and the unlocking reaction that we use on our A and B rings. Cleavage of a plain C_{66} ring and of a C_{66} ring with a knot in it would lead to topologically identical products (because an unknotted string is topologically equivalent to a knotted one).

The trefoil knot possesses one property that single rings do not: it can exist in two forms. One is a "right-hand" knot and the other a "left-hand" knot. They are mirror images of each other and have the same sort of relation as right and left hands. Left-handed and right-handed molecules are optically active: in solution they rotate the plane of polarized light in opposite directions. This rotation of light should serve as a tracer for the knot.

Unfortunately our solution of C_{66} rings, including the knotted rings, will contain equal numbers of left and right forms and will produce no net rotation of polarized light. We hope to separate out at least some of one type of knot by introducing another optically active compound that will absorb or hold one of the knots rather than the other (somewhat as a right hand clasps a left hand but not another right). Then we may be able to separate one type from the mixture by chromatography and to show its presence through the use of polarized light.

Some circular compounds are found in living organisms and may exhibit topological isomerism. These rings, how-

ever, do not arise at random. They are synthesized in close proximity to the surface of an enzyme and are unlikely to be interlocked or knotted unless nature "intends" them to be so. One unusual material is a virus deoxyribonucleic acid (DNA) molecule that appears to be circular [see "Single-stranded DNA," by Robert L. Sinsheimer; *SCIENTIFIC AMERICAN* Offprint 128]. The ring contains 5,500 carbon-oxygen-phosphorus-oxygen groups, for a total of 22,000 atoms in the loop itself. The possibilities for knotting are great. The natural ring may even have one or more knots, but there is no evidence to support this speculation.

Topological chemistry is not restricted to the single-stranded molecules considered so far. Parallel strings of atoms can be cross-linked to produce a "surface," which can then be transformed in many

ways. The first example has already been prepared by John F. Brown, Jr., and his co-workers at the General Electric Research Laboratory. They have synthesized a double-stranded silicone. (In silicones the atoms of the main chain are silicon and oxygen, not carbon.) A particularly interesting possibility is the conversion of a roughly rectangular surface into a Möbius strip by a half-twist about the long axis before the ends are joined to form a ring. Subsequent cleavage of the cross-links in such a two-stranded strip would lead to a large single ring, since there is only one edge to the surface [see *bottom illustration on page 748*]. A three-stranded strip would produce a pair of interlocked rings. Knots may be formed also: separation of a two-stranded strip that has three half-twists will yield a trefoil.

Whether or not catenanes and knots

will have any properties of special interest can only be determined after considerably more experiment. One possibility, suggested by Herman F. Mark of the Polytechnic Institute of Brooklyn, is based on the preference of some inorganic compounds for existing in small ring forms instead of linear structures of high molecular weight. To take advantage of their stability at high temperatures and to obtain long polymeric structures, similar to those now found only among carbon compounds, many loops might be interlinked. The statistical methods that we have employed for our catenanes would not be suitable because of their small yield; other methods might be developed if the properties of the simplest inorganic catenanes were promising.

The Author

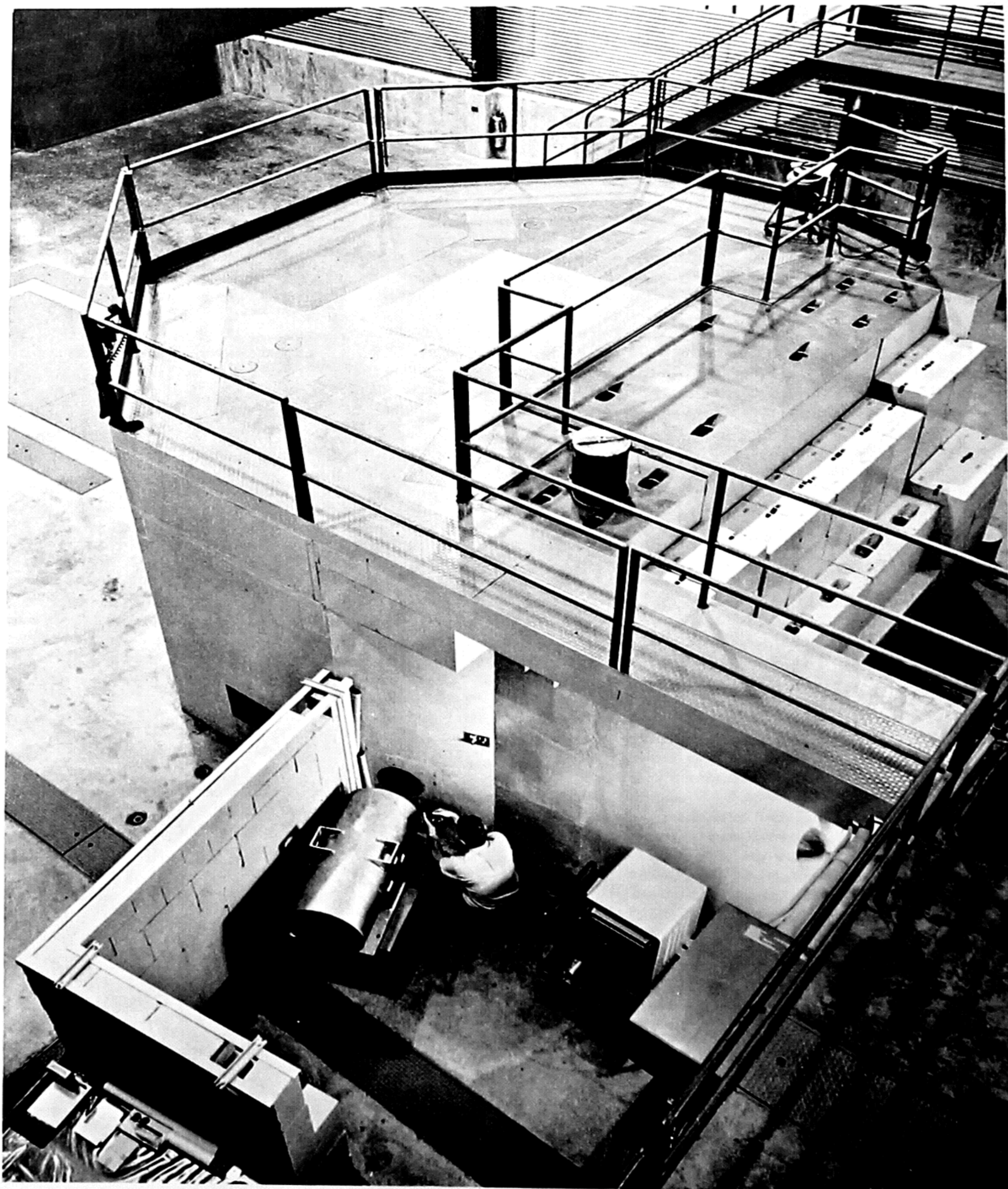
EDEL WASSERMAN, a member of the technical staff of the Bell Telephone Laboratories, is spending the current academic year as visiting professor of chemistry at Cornell University. Born in New York City, Wasserman took his B.A. at Cornell in 1953 and then did graduate work at Harvard University, where he studied under the late William Moffitt. He received his Ph.D. in chemistry in 1959. Wasserman joined the chemistry department of the Bell Laboratories in 1957.

Bibliography

CHEMICAL TOPOLOGY. H. L. Frisch and E. Wasserman in *Journal of the American Chemical Society*, Vol. 83, No. 18, pages 3789-3795; September 20, 1961.

THE PREPARATION OF INTERLOCKING RINGS: A CATENANE. E. Wasserman in *Journal of the American Chemical Society*, Vol. 82, No. 16, pages 4433-4434; August 20, 1960.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound *SCIENTIFIC AMERICAN* Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.



"JUGGERNAUT" RESEARCH REACTOR at the Argonne National Laboratory provides an intense beam of neutrons useful for radiography. The beam leaves the reactor through a slot opposite

the horizontal cylinder. The object to be radiographed is mounted in front of the cylinder, which stops the beam. A close-up of the cylinder and neutron exit port appears at the top of page 755.

NEUTRON RADIOGRAPHY

by Harold Berger

Pictures made by slow neutrons disclose many things that cannot easily be seen in pictures made by X rays. New detection methods are making it attractive to use neutrons as an inspection tool.

The value of radiography in medicine and industry is so firmly established that today the amount of photosensitive film used for making X-ray pictures greatly exceeds the amount of black-and-white film used by amateur photographers. With X rays of sufficient energy and intensity it is possible to make pictures through objects of any material and of almost any thickness. It has been known virtually from the time the neutron was discovered, 30 years ago, that neutrons might also be used to make pictures through objects. But since neutrons in a form useful for imaging purposes are harder to produce than X rays, one may wonder what advantage is to be gained by making neu-

tron radiographs. The answer is that neutrons make it possible to see things that X rays do not reveal at all, or reveal only with difficulty.

Neutron radiography has become practical with the advent of nuclear reactors and particle accelerators, which can provide a source of neutrons of the required intensity. Nuclear reactors produce neutrons directly, as a by-product of nuclear fission. Accelerators produce beams of high-energy charged particles, usually protons, which can give rise to neutrons if they are directed against a suitable target material, such as lithium or beryllium. The high-energy neutrons produced by an accelerator are absorbed about equally by different materials. To

be most useful in radiography neutrons must be slowed down until their velocity is simply the velocity imparted to them by the random movement of atoms at ordinary temperatures. Such low-energy neutrons, called thermal neutrons, can be obtained directly from nuclear reactors.

The value of thermal neutrons in radiography is that their absorption characteristics are quite different from those of X rays. For elements at either end of the periodic table the absorption characteristics of thermal neutrons and X rays are essentially reversed [*see illustration on next two pages*]. Heavy elements such as lead, bismuth and uranium are practically transparent to thermal neutrons,



NEUTRON RADIOGRAPH OF GRASSHOPPER gives information that would be difficult to obtain in an X-ray picture. Tissue com-

ponents absorb X rays about equally. Neutrons, however, are heavily absorbed by hydrogen. Thus white areas here are hydrogen-rich.

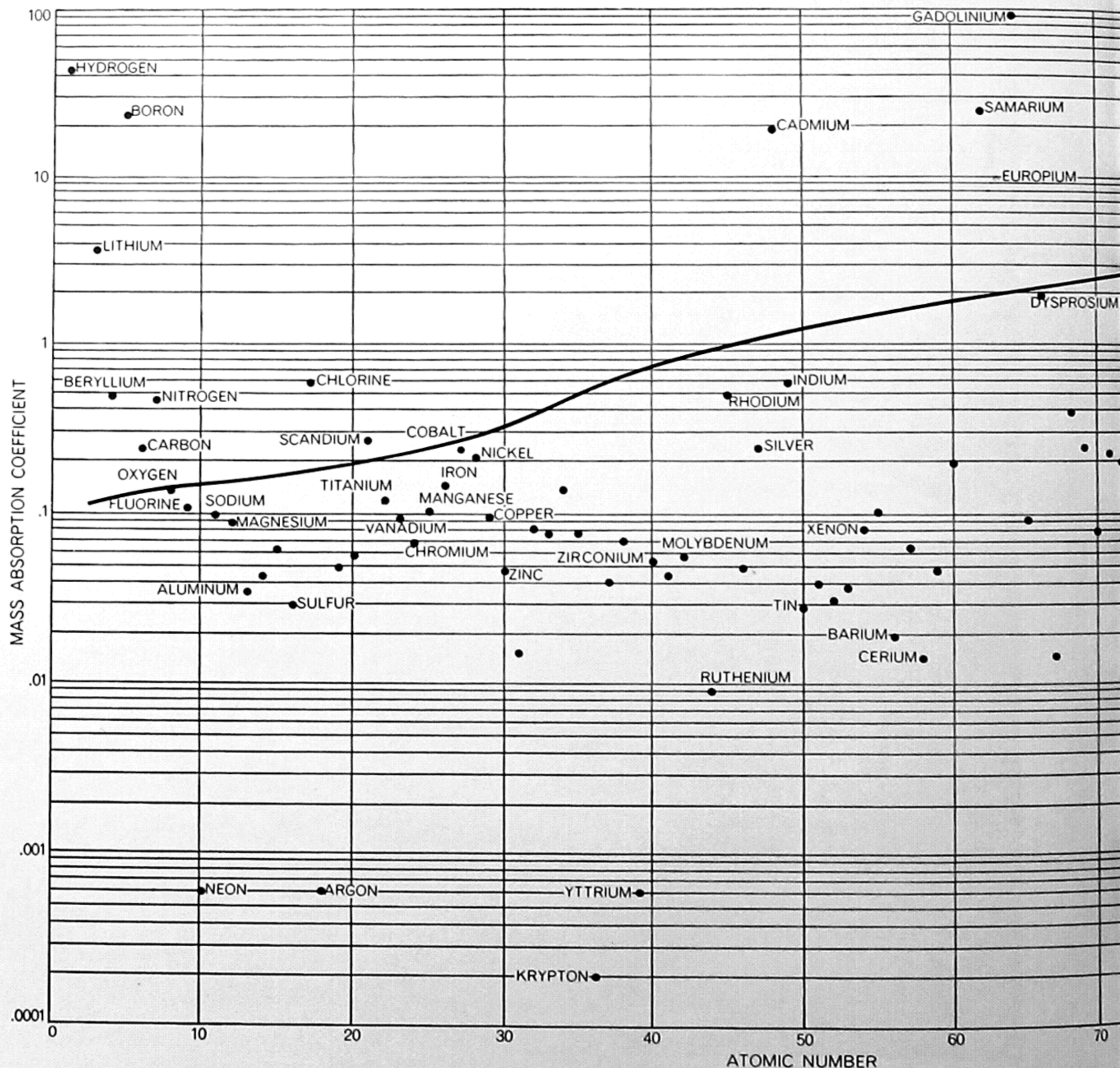
whereas they absorb X radiation strongly. Conversely, hydrogen, lithium, boron and other light elements strongly absorb thermal neutrons but allow X rays to pass freely. For example, with neutron radiography it would be an easy matter to record the height of a column of water in a lead tube. Neutrons would have no difficulty penetrating the lead but would be strongly absorbed by the hydrogen in the water. X rays would be so readily

absorbed by the lead that the slight additional absorption due to water would be difficult to observe.

In general, various materials stop X rays in direct proportion to their mass. For this reason it is difficult to distinguish between elements that lie near each other on the periodic table. Neutron absorption, on the other hand, is not related directly to atomic number or mass; neighboring elements such as cad-

mium and tin or boron and carbon can differ in neutron "transparency" by factors of 100 or 1,000.

A neutron radiograph is made in much the same way as an X radiograph. The object to be inspected is simply placed between a source of radiation and a radiation detector. If the source produces fast neutrons, they can be slowed down by allowing them to diffuse through a medium containing light elements, such



ABSORPTION CAPACITY OF ELEMENTS is plotted for X rays (black curve) and for thermal, or low-energy, neutrons (colored dots). The higher the mass-absorption coefficient, the more absorb-

ing the material. In general the coefficient for X rays increases steadily with atomic number. The coefficient for neutrons is much more random, but there is a tendency for absorption to be the

as hydrogen or carbon. Since a great many fast neutrons are absorbed in the thermalizing process, the original source must produce on the order of 10 billion neutrons per second to yield a thermal-neutron beam of useful intensity. We have found at the Argonne National Laboratory that an intensity of at least 100,000 thermal neutrons per square centimeter per second appears desirable. Even low-power reactors, however, can

produce beams 100 times more intense.

A number of commercially available accelerators can produce thermal-neutron beams of adequate intensity at a price competitive with high-energy X-ray equipment. Recently Edward J. Hennelly of the Savannah River Laboratory operated by E. I. du Pont de Nemours and Company has described a low-cost radioactive source of thermal neutrons. The source consists of radioactive antimony 124, which has a half life of 60 days, and beryllium. Neutrons are produced when the beryllium atoms are struck by gamma rays from the antimony 124. The initial cost of the source has been estimated at \$3,500, and the cost of maintaining the source at a high output should be less than \$5,000 a year.

One problem associated with fast-neutron sources is that the thermalized neutron beam also contains high-intensity gamma radiation. If the neutron beam were being used to inspect uranium or some other heavy metal, the gamma rays might even contribute usefully to the image. In other cases, however, gamma radiation may confuse the radiographic picture. One answer to the problem is to remove the gamma rays by means of a filter, consisting of an element such as bismuth, that has high absorption for gamma radiation but low absorption for thermal neutrons. Another approach is to allow the neutrons to strike a crystal and form a reflected beam that contains thermal neutrons but little or no gamma radiation. Such an arrangement is called a neutron spectrometer; it not only reduces gamma rays but also fractionates the neutrons into beams of different energies, or wavelengths. By selecting only neutrons reflected at a certain angle one can obtain a "monochromatic" beam in which all the neutrons have about the same energy. It happens that most of our radiographic studies at Argonne have been done with monochromatic beams, but such beams offer no advantage for most applications.

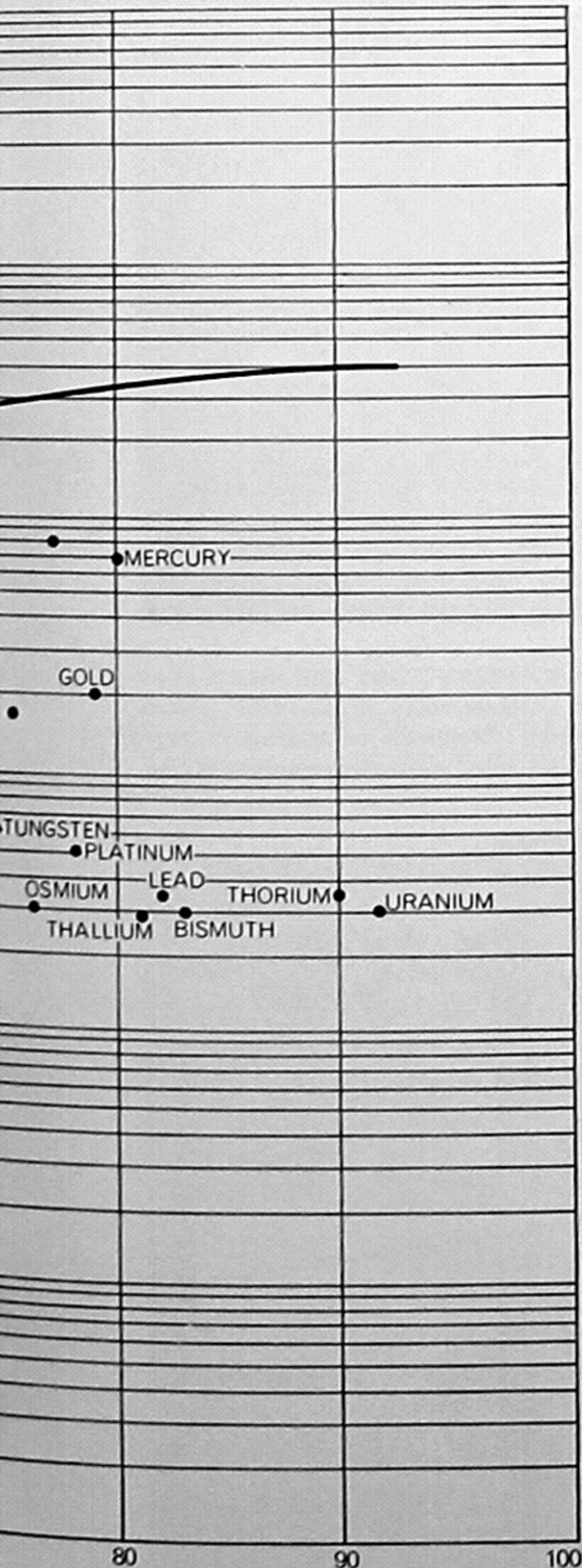
Still another way to combat the problem of gamma-ray contamination is to use detectors that do not respond to gamma rays. I shall therefore broadly describe the various methods of detecting neutrons. In the transfer method the neutrons are allowed to strike a metal screen that becomes radioactive in proportion to the intensity of the neutrons at each point. The screen thus contains a radioactive image of the test object. To make this image visible the screen is placed against photographic film, and the radiation arising from radioactive decay exposes the film. Since the photographic film is not exposed to the neutron

beam itself, the gamma rays in the neutron beam are made ineffective.

Metal screens for the transfer method incorporate isotopes that have a conveniently short half life. Examples are gold 198 (half life, 2.7 days), dysprosium 165 (2.3 hours), indium 116 (54 minutes), rhodium 104 (4.4 minutes) and silver 108 (2.3 minutes). Even with saturation exposure the Argonne neutron beam with an intensity of 100,000 neutrons per square centimeter per second yields a barely detectable transfer image when short-lived rhodium and silver are used in conjunction with a fast X-ray film. The image cannot be strengthened by lengthening the exposure because after a certain time—equal to about three half lives of the isotope in the screen—the number of atoms decaying begin to match the number of new radioactive atoms being produced. To achieve a stronger image with very short-lived isotopes it is necessary to increase the intensity of the neutron beam. As it is increased more radioactive atoms are created per unit of exposure time, and the level of radioactivity in the screen is increased correspondingly. As a result rhodium and silver isotopes, as well as many others, can be used to make transfer radiographs with beams that have an intensity greater than 100,000 neutrons per square centimeter per second.

Another neutron-detection method can employ a wide range of screen materials, including some that do not actually become radioactive; it also responds, however, to the gamma radiation in the beam. In this technique, which we have termed the direct-exposure method, the screen and the film are exposed to the neutron beam together. The neutrons have little direct effect on the film, but the film now accumulates all the radiation emitted from the screen during the exposure. In addition to collecting all the radioactive-decay radiation, the screen also collects "prompt emission" radiation, which can consist of either gamma rays or charged particles (usually alpha particles, or helium nuclei) produced almost instantaneously by neutron bombardment. This prompt emission combines with the decay emission to allow much shorter exposures with the direct-exposure method than with the transfer method.

Several materials that have only a negligible tendency to become radioactive make excellent direct-exposure screens. These materials include cadmium and gadolinium, which are prompt gamma-emitters, and certain isotopes of boron and lithium that emit alpha particles when they are struck by slow neu-



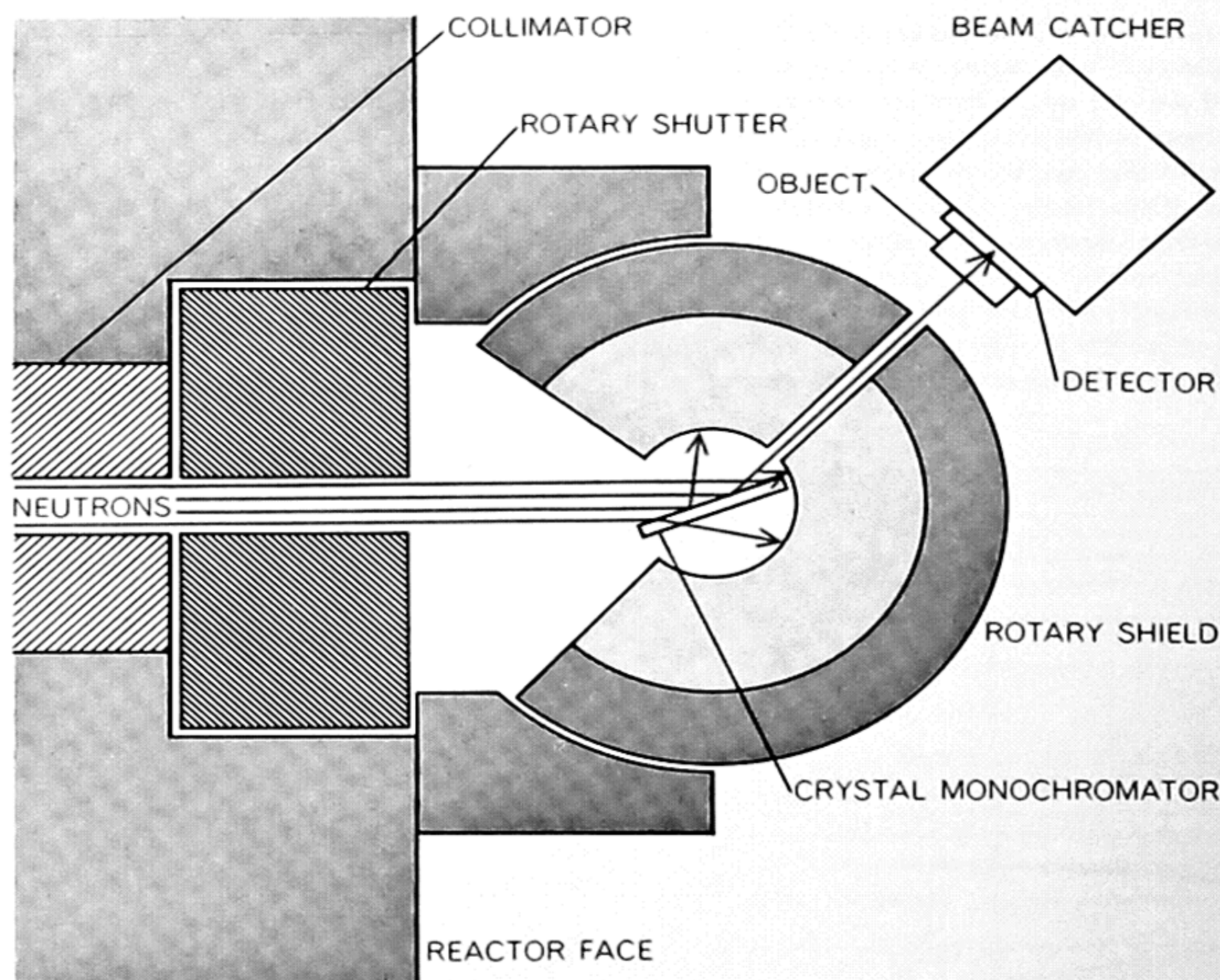
reverse of that for X rays. Values are based on those given by James Thewlis of the Atomic Research Establishment at Harwell.

trons. The most sensitive detectors are made by combining the alpha-emitters boron 10 and lithium 6 with a phosphor, such as silver-activated zinc sulfide, which gives off flashes of light when it is struck by alpha particles. These scintillations expose the film more effectively than alpha particles themselves do. Such detectors produce good images when struck by only a few hundred neutrons per square millimeter. This is at least an order of magnitude more sensitive than any other neutron detector that seems useful for radiography.

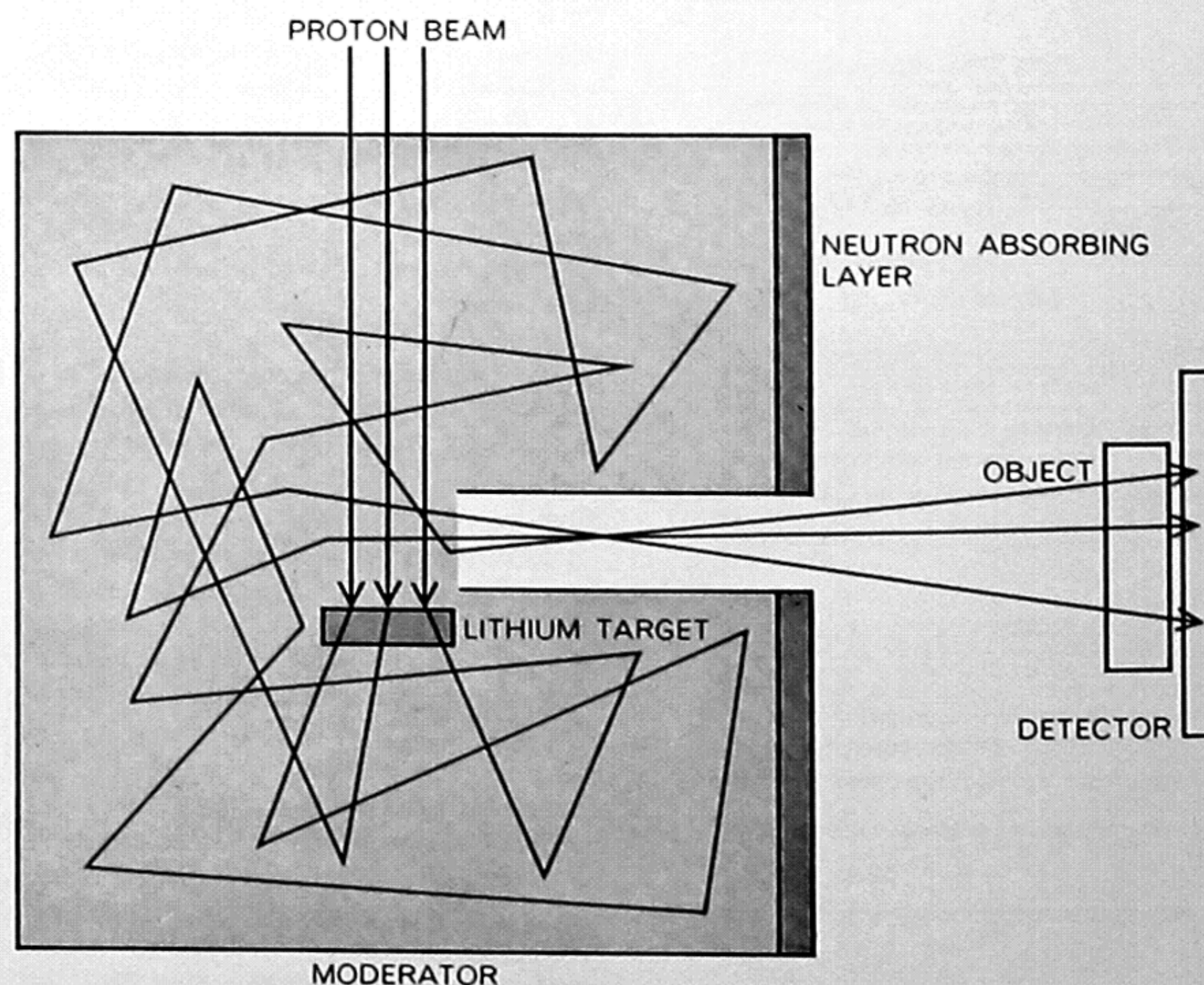
The over-all sensitivity of the detector depends, of course, on the speed of the film employed. Until recently the shortest exposures were obtained with scintillators placed in direct contact with blue-sensitive X-ray films, such as Kodak Type F, or with the Polaroid film that has a speed rating of 3,000. The newly available Polaroid film with a rating of 10,000 has cut exposure times still further. It is possible to detect neutrons directly in ordinary photographic film, without the need for screens, and one can also use special emulsions loaded with boron 10 or other elements that produce radiation when struck by neutrons, but the speeds are still below those obtainable by converter-screen methods.

The various detection methods differ not only in speed but also in image sharpness. To examine sharpness, or resolving power, we have made many neutron radiographs of test objects containing small holes. We find the highest resolution is obtained with thin gadolinium screens in a direct-exposure method. Test films made by this method produced distinct images of tiny holes that are appreciably closer than 1,000th of an inch. The sharpness can evidently be attributed to the fact that gadolinium, when struck by thermal neutrons, emits low-energy electrons that travel only a short distance in the photographic emulsion. Next in order of image sharpness are most of the transfer methods and the direct exposures using scintillators. Other direct-exposure methods, using screens of such materials as cadmium, silver and indium, yield somewhat poorer images.

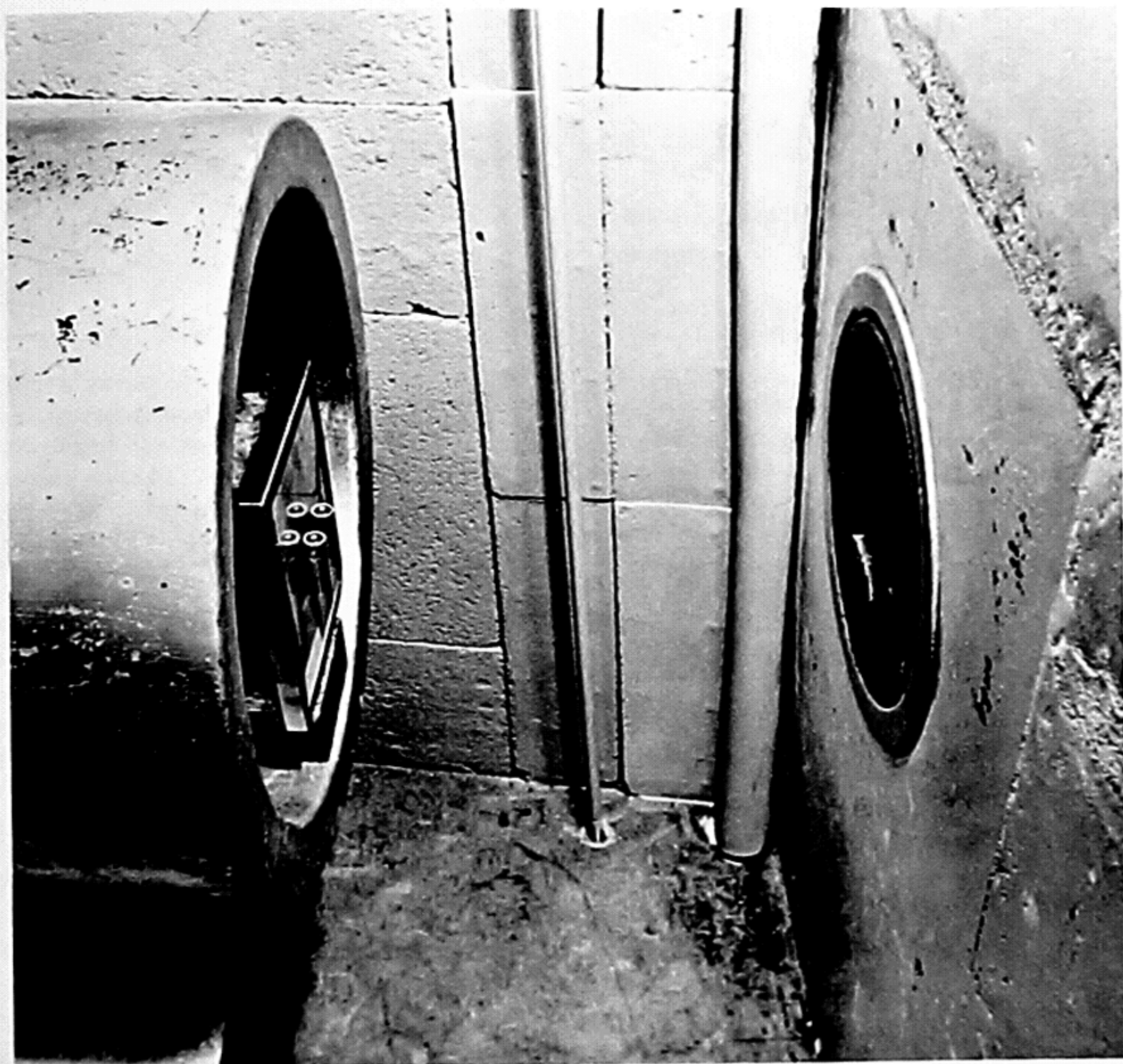
Another important property of a radiographic image is contrast, which makes it possible to distinguish different thicknesses of material in the test object. In examining the heavy metals, for example, we find that we can usually detect thickness variations of the order of 1 per cent with any of the detection methods, except when a scintillator is employed. In



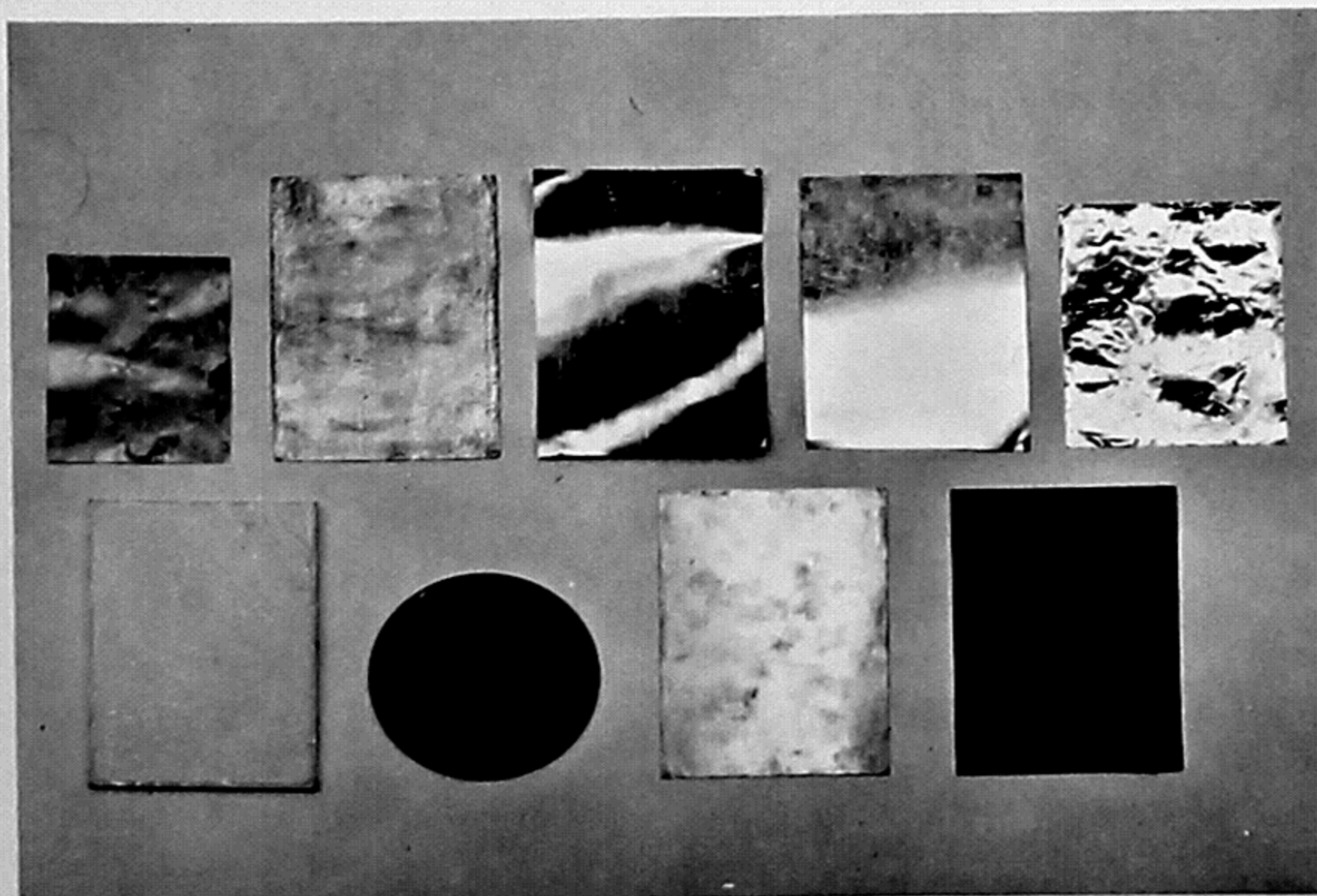
REACTOR NEUTRON SOURCE is of the type used by the author in experiments at the Argonne National Laboratory. After leaving the reactor (*left*) the neutrons strike a large single crystal, where they are deflected at various angles depending on their energies. For most radiographic purposes the neutron beam could be used just as it leaves the reactor.



ACCELERATOR NEUTRON SOURCE yields high-speed neutrons when high-energy protons strike lithium or some other suitable target. The neutrons are slowed down by diffusing through a moderator, which contains light elements such as hydrogen or carbon.



SETUP FOR NEUTRON RADIOGRAPHY is demonstrated at the Argonne National Laboratory. The neutron beam emerges from the Juggernaut reactor through the port at the right. Objects to be examined (small dry batteries) are mounted on a detector screen. Neutron radiographs of the batteries are compared with X radiographs below on page 757.



RADIOGRAPHIC SCREENS convert neutron radiation into radiation that exposes photographic film. When struck by neutrons, the screens in the top row (dysprosium, indium, rhodium, silver and gold) become radioactive; the two screens at right in the bottom row (cadmium, gadolinium) emit gamma rays; the other two (boron scintillators) give off light.

images produced by a scintillator the smallest detectable thickness variation runs from about 5 to 15 per cent. Although the scintillator methods leave something to be desired, we have found them useful for a quick look at a new test object. Speed and convenience are particularly good when the scintillator image is recorded on the high-speed Polaroid film, which develops in only 10 seconds.

Generally speaking, the resolution and contrast of the better radiographic methods compare favorably with those commonly obtained with conventional X-ray and gamma-ray equipment, but they do not yet measure up to the best results obtainable with X rays. Nevertheless, the quality of present neutron radiographic methods is high enough to suggest many industrial and technical applications, particularly where neutrons can provide distinctive information.

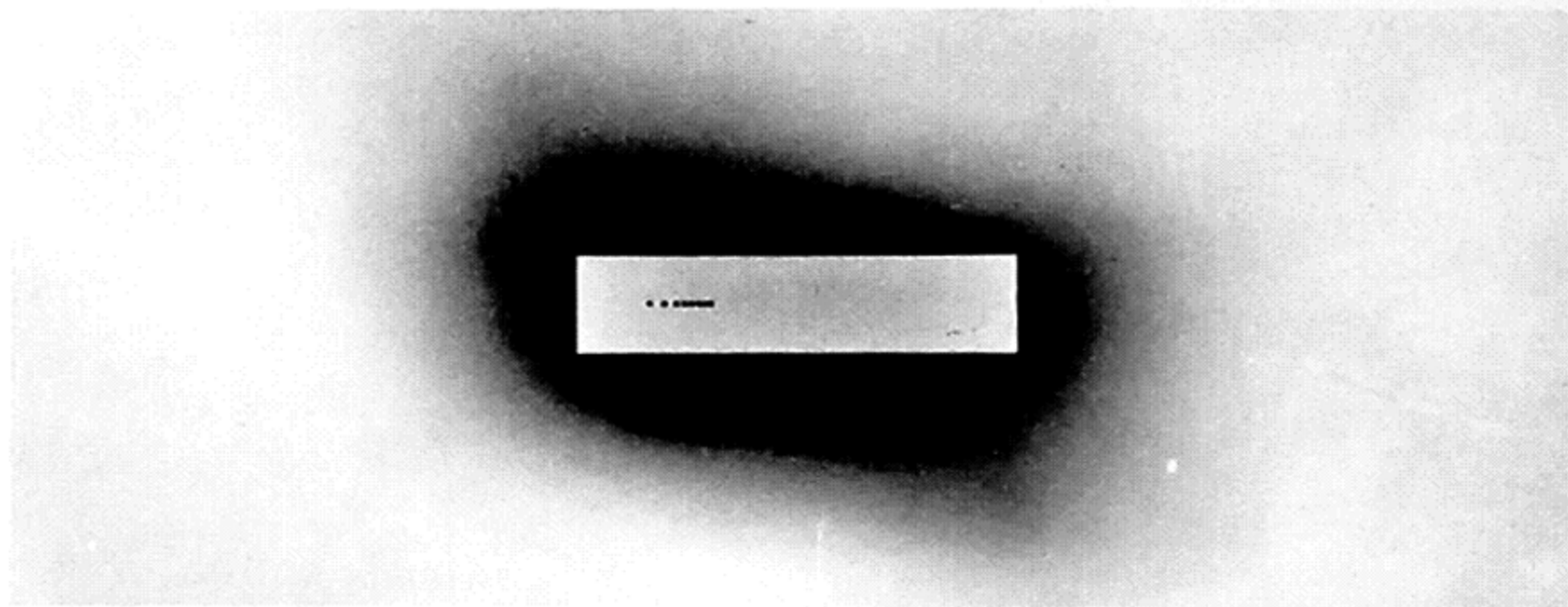
In metallurgy, for example, neutron radiography could be used to observe concentrations of light elements such as hydrogen, lithium or boron contained within various kinds of materials or objects. The top illustration on page 757 is a neutron radiograph of a zirconium bar containing sintered boron carbide, used to control the neutron intensity in portions of a nuclear reactor. A critical component in the bar is boron, placed there because of its capacity for absorbing neutrons. One would like to have an inspection method that would show whether or not the boron is distributed uniformly. An X radiograph is not much help because boron carbide and zirconium absorb X rays about equally. Neutrons, however, are absorbed several hundred times more readily by boron than they are by zirconium; thus the distribution of the boron shows up clearly in a neutron radiograph. Going through the periodic table, one can find many other combinations of substances that are easily distinguished by neutron absorption but not by X-ray absorption.

Another metallurgical use for neutron radiography is in the inspection of objects containing heavy metals such as uranium, lead and bismuth, which are much more transparent to neutrons than they are to X rays. When the thickness of such metals runs beyond a few inches, the exposure time for a neutron radiograph can become significantly less than for an X radiograph [see illustration on page 756]. Even the low-intensity neutron beam at Argonne can be used, with high-resolution techniques, to inspect a sample of uranium two inches thick in about an hour. This is only a small fraction of the comparable exposure that

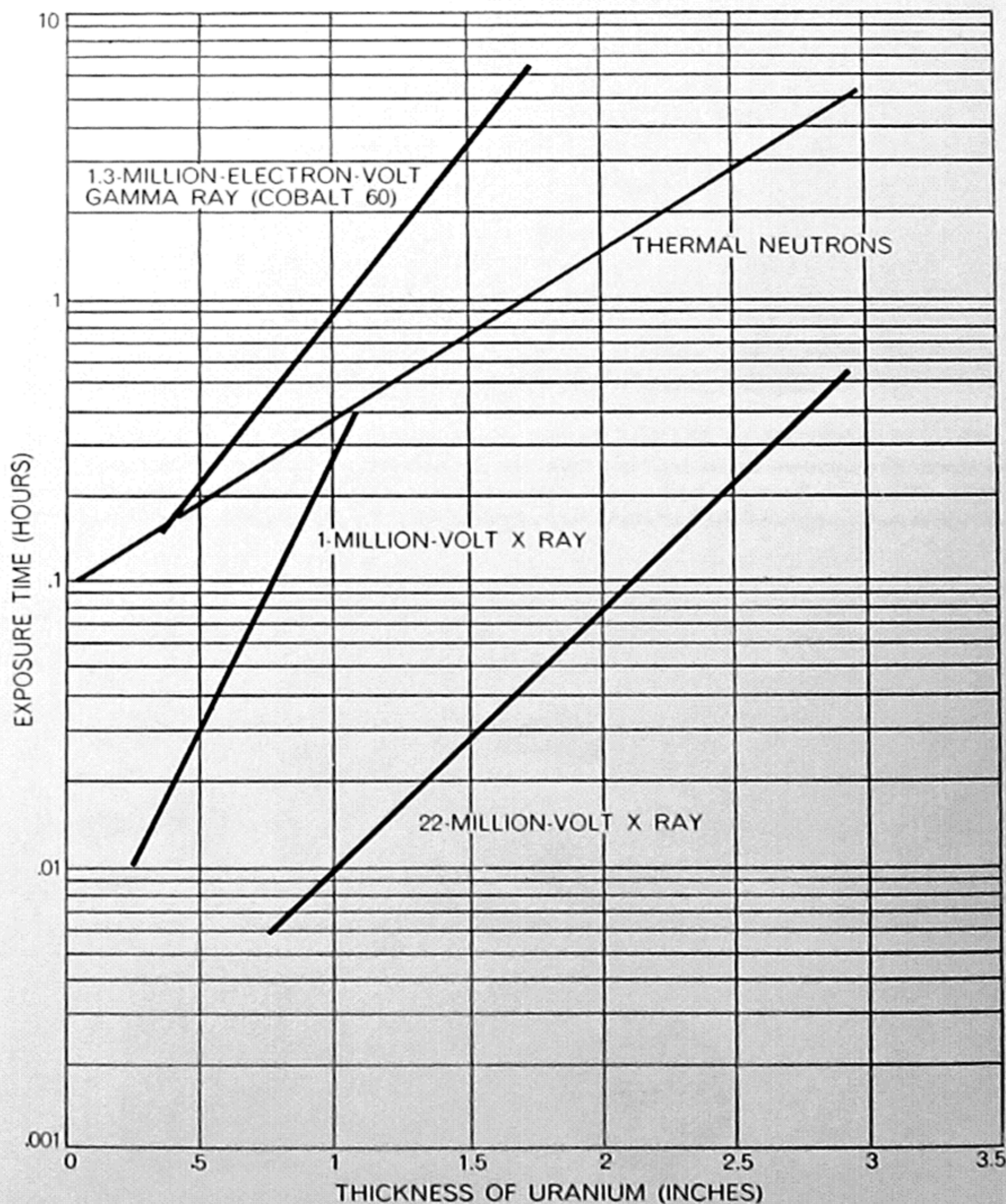
would be required with a standard one-million-volt X-ray generator or with gamma rays from a cobalt-60 source of fair intensity. With neutron beams of greater intensity than our original Argonne beam, which should not be difficult to achieve, the exposure time could be cut to less than a minute. In comparison, the 22-billion-volt X rays from a betatron would require several minutes to produce a comparable radiograph through two inches of uranium. Actually with the fast scintillator technique the Argonne beam could produce a neutron radiograph of such a sample in less than 30 seconds, but the quality of the picture, as I have explained, would be substantially below X-ray quality.

One of the fascinating things about thermal neutrons is that the same neutron beam used to inspect several inches of uranium or lead can be used to inspect specimens such as leaves, insects and thinner biological specimens. The neutron picture of a grasshopper on page 751 shows the possibility of applying neutron radiography in biological studies. I do not mean to suggest that neutron radiography of living animals or humans is feasible, much less desirable. Quite the contrary; the rich hydrogen content of animal tissue would make it difficult for the neutron beam to penetrate, and the exposures needed might induce enough radioactivity in the tissue to harm the organism. Neutron inspection of thin biological specimens may nonetheless offer advantages, particularly if the purpose of the inspection is to locate the position of hydrogen or other neutron-absorbing materials. For such biological applications the exposures required for neutron radiography are not excessive. With the use of metal screens and fast X-ray films, a radiograph of reasonable quality can be obtained with a total exposure of 10 million thermal neutrons per square centimeter of detector. This can be translated into an exposure of about 10 milliroentgens, a figure comparable to that required by fast X-radiographic methods.

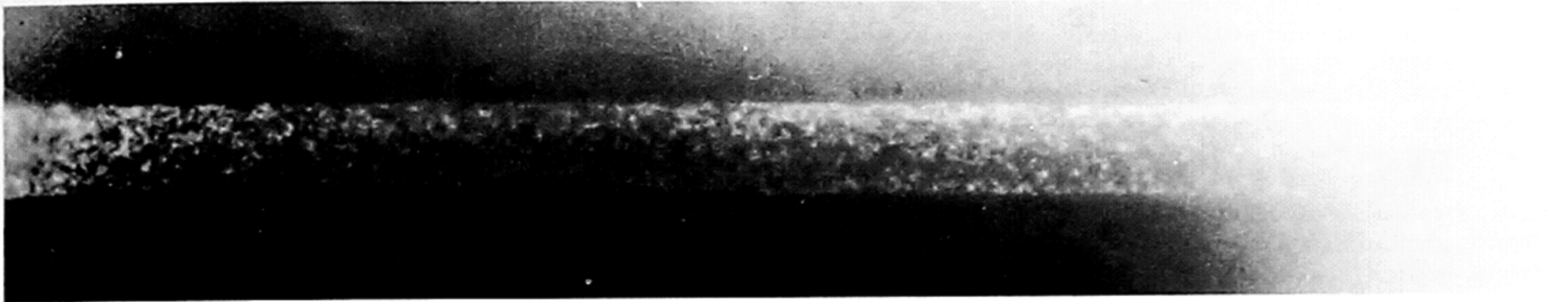
The strong neutron absorption of hydrogen also makes inviting the possibility of neutron-radiographic inspection of such hydrogen-containing materials as paper, rubber, wood, plastics and adhesives. Small differences in the hydrogen content or the thickness of such substances, as well as the location of the substances in objects primarily made up of other materials, should be comparatively easy to observe with neutron radiography. The radiographs of the small batteries at the bottom of page 757 show



RESOLUTION TEST shows that neutron radiography can record fine detail. The test object is a piece of cadmium containing eight holes .02 inch in diameter. The picture was made with a thin gadolinium screen by the direct-exposure method described in text. The irregular blackened area around the object shows the actual dimensions of the neutron beam.



RADIOGRAPHIC EXPOSURE TIMES are shown for neutrons, for gamma rays from cobalt 60 and for X rays of two levels of energy. Gamma rays and X rays, of course, are both high-energy photons and carry different names only by custom. Exposure times reflect typical values for X-ray and gamma-ray sources and are based on the work of Gerold H. Tenney of the Los Alamos Scientific Laboratory. Neutron exposures are for an intensity of 100,000 neutrons per square centimeter per second, using a gadolinium screen and direct exposure.



ZIRCONIUM BAR containing boron carbide makes a good test object for neutron radiography. Since boron carbide and zirconium have about the same absorption for X rays, an X radiograph would

not tell much about boron carbide distribution. Boron, however, is one of the strongest absorbers of neutrons, whereas zirconium is only an average absorber. Hence boron carbide shows up as white.

the different responses of neutrons and X rays to objects containing various substances, including hydrogen. The pictures are X radiographs and neutron radiographs of new and used batteries. One can see first of all that the waxy material in the top seal of the battery—a material rich in hydrogen—completely absorbs the neutrons but transmits much of the X radiation.

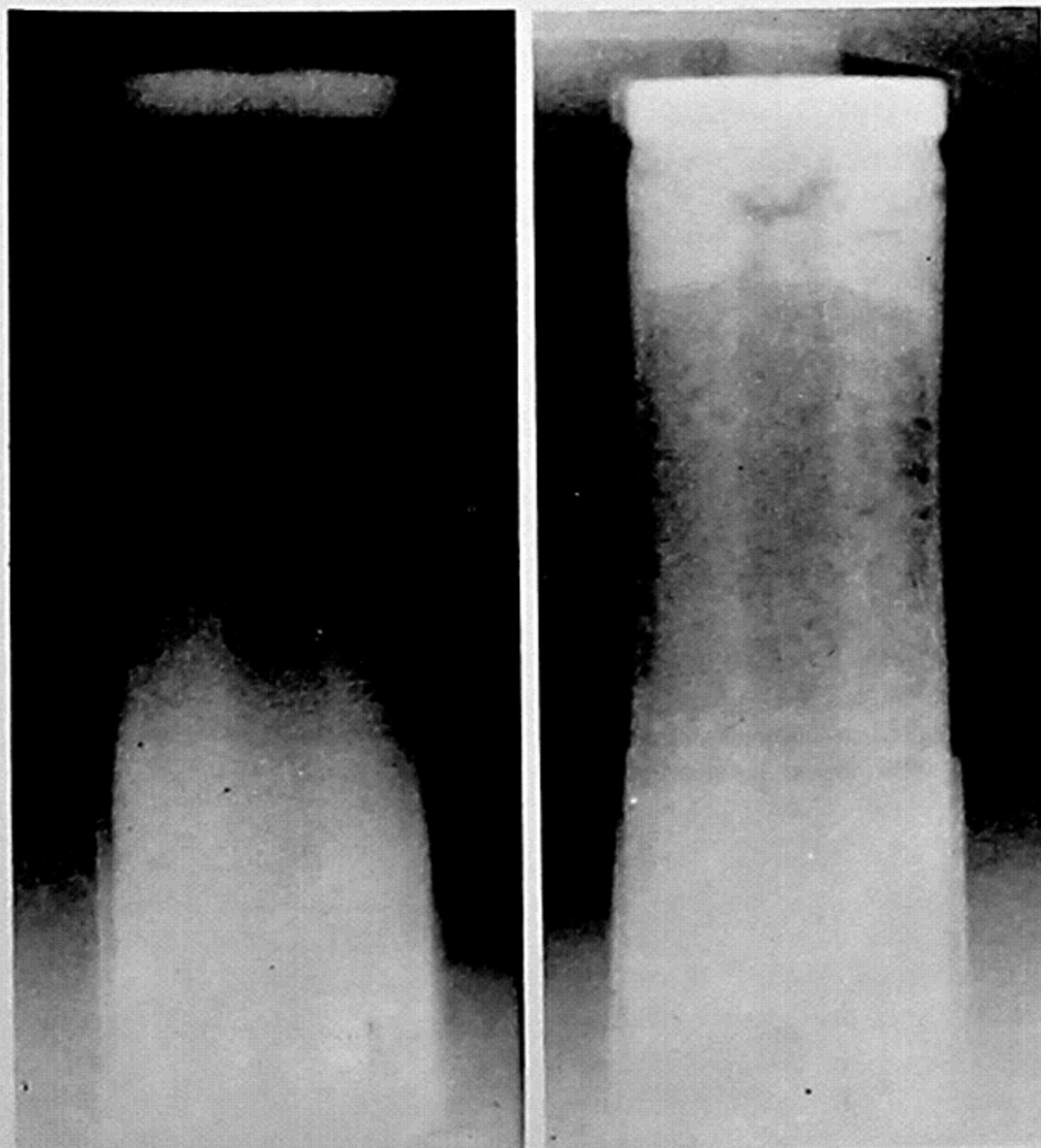
A second and more interesting difference in response can be found if one compares the new and the used batteries. This difference is most evident in the expansion chamber near the top of the battery. In a new battery the expansion chamber is empty; both neutrons and X

rays pass through it easily. In a used battery the expansion chamber has become filled with a paste-like material of high hydrogen content. The material consequently absorbs neutrons strongly but for X rays is only slightly more absorbent than empty space.

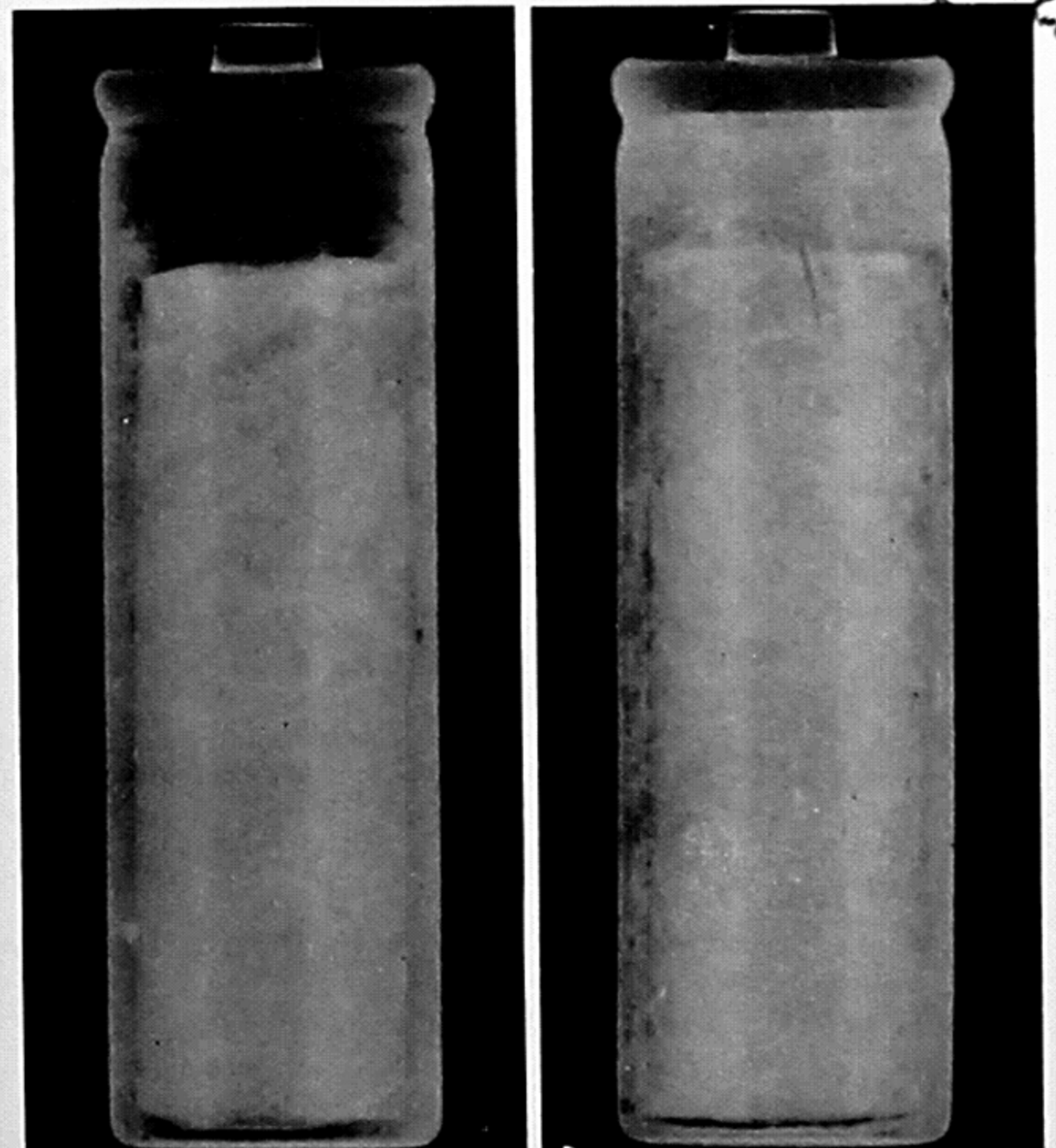
These neutron and X radiographs of the same samples demonstrate how the two techniques complement each other and how they broaden the usefulness of radiographic inspection when they are used together. Still, in spite of the many potential advantages to be gained using neutron radiography, and in spite of the fact that some of the original work on neutron radiography was reported in the

technical literature more than 15 years ago, little use has been made of this inspection method to date.

This is perhaps not too surprising; the availability of neutron sources that have characteristics useful for radiography is still not very great. It is also true that until recently the quality of inspection possible with neutron radiography was not widely known, nor was much known about the many useful detection methods that could be employed for neutron imaging. As a result anyone wishing to try neutron radiography for a special inspection problem had the task of finding suitable equipment and of working out his own exposure methods. Even now,



RADIOGRAPHS OF SMALL DRY BATTERIES were made with neutrons (pair at left) and X rays (pair at right). In each set of pictures the battery on the right is a used one, hence the expan-



sion chamber near the top is filled with an ammonia-containing material. This material produces a strong increase in neutron absorption but only a modest increase in X-ray absorption.

U.S. Patent Collection

with many of these difficulties much diminished, there are still problems standing in the way of the routine application of neutron radiography.

In order to inspect large objects, or to view many small objects in one exposure, it would be desirable to have a larger beam of radiation than any now available. Most of our experiments were done with a neutron beam that covered an area only about three inches in diameter. The beam from the recently completed Juggernaut reactor at Argonne, shown on page 750, will produce radiographs 2.5 by four inches and will eventually sup-

ply a neutron beam six inches in diameter. It is possible, of course, to radiograph large objects section by section with small beams, but obviously this is slow and cumbersome. In short, the difficulty of obtaining large, uniform neutron beams of high intensity is still a major deterrent to the widespread application of neutron radiography.

For many purposes, particularly scientific ones, it would also be valuable to have a well-defined diverging beam of neutrons. Such a beam would in effect constitute a neutron microscope. It could be used, for example, to observe hydride

precipitation in single crystals of metal or to observe the diffusion of boron or lithium in semiconductor materials. A diverging neutron beam would also have many uses in biology.

Better neutron sources are therefore needed to help neutron radiography realize its full potential. Even now, however, the capabilities and potential advantages of neutron radiography recommend its consideration for a variety of inspection problems. There is not much doubt that it will eventually become a routine inspection tool.

The Author

HAROLD BERGER is associate physicist with the Nondestructive Testing Group of the Metallurgy Division at the Argonne National Laboratory. Berger studied physics at Syracuse University, receiving B.S. and M.S. degrees there in 1949 and 1951 respectively. From 1950 to 1959 Berger was a physicist in the Advanced Development Laboratory of the General Electric Company in Milwaukee, Wis. After a brief period as senior physicist with the Solid State Devices Division of the Battelle Memorial Institute, Berger went to Argonne in 1960.

Bibliography

- COMPARISON OF SEVERAL METHODS FOR THE PHOTOGRAPHIC DETECTION OF THERMAL NEUTRON IMAGES. Harold Berger in *Journal of Applied Physics*, Vol. 33, No. 1, pages 48-55; January, 1962.
- A DISCUSSION OF NEUTRON RADIOGRAPHY. Harold Berger in *Nondestructive Testing*, Vol. 20, No. 3, pages 185-194; May-June, 1962.
- NEUTRON OPTICS. D. J. Hughes. Interscience Publishers, Inc., 1954.
- NEUTRON RADIOGRAPHY. H. Kallmann in *Research*, Vol. 1, No. 6, pages 254-260; March, 1948.
- NEUTRON RADIOGRAPHY. J. Thewlis in

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

THE CONDUCTION OF HEAT IN SOLIDS

by Robert L. Sproull

It is a familiar observation that some materials are better conductors of heat than others. The process is governed by the properties of a unit of sound energy known as the phonon.

A sterling silver spoon and a silver-plated spoon immersed in a cup of hot coffee serve to demonstrate that solid materials conduct heat and that some conduct heat better than others. Early in the history of technology it became necessary to measure this property in materials of construction. Such measurements were made with great precision in the absence of any theory to explain the conduction of heat. They showed, for example, that metals are better heat conductors than non-metals long before it was suspected that metals are also good conductors of something called electricity. And measurements of both properties in various materials were made before it occurred to anyone to connect these properties with the fact that solids also conduct sound. Today the demands of technology and fundamental questions put by science have carried the measurement of heat conduction by both metals and nonmetals into the extreme ranges of the temperature scale. From such work has come an understanding of this property of matter that relates it to the conduction of both sound and electricity. The conduction of heat can now be described in terms that take account not only of the particulate structure of matter but also of the particulate nature of energy, as expressed in the powerful generalizations of quantum theory.

Although many elaborate and ingenious methods for measuring the thermal conductivity of solids have been invented, the simplest method is still commonly used. One end of a bar of the solid is heated, perhaps by an attached electrical heater; the power input is measured, and the temperature difference between two points on the bar a measured distance apart is also measured. The thermal conductivity—the heat flow per unit

of temperature difference in a standard length and cross section of a specimen—can then be calculated from these data by correcting for the size and shape of the sample. In a given material the heat flow is obviously larger through a short, thick specimen for the same temperature difference than it is through a long, thin one.

The finding that metals exhibit high electrical conductivity as well as high heat conductivity yielded the first insight into the primary mechanism of heat conduction in those materials. Metals owe their high heat conductivity to their abundance of free electrons, the familiar carriers of electricity. If one end of a bar of a metal is heated to a higher temperature than the other, the electrons in the hotter end acquire faster speeds than those in the cooler end do. Since the electrons in a metal belong to the solid as a whole and not to individual atoms, the energetic electrons can flow to the cold end and the less energetic electrons can flow to the hot end. Thus kinetic energy is transferred to the cold end, warming it.

What characteristics of a particular metal determine whether the conduction of heat will be greater or less than in another metal? Theory and a great deal of experimental evidence show that the heat flow is greater the farther each electron can move before it is diverted. If an energetic electron could move throughout a metal specimen without being scattered or losing its energy, the metal would have an infinite thermal conductivity. It might at first be thought that an electron would be constantly colliding with the closely packed atoms making up a solid. In a perfectly organized crystal, however, in which the atoms are all placed in an ex-

actly regular array in the crystal lattice, an electron would travel a path of infinite length before being scattered. An electron cannot give up any of its kinetic energy to such a crystal. The perfect solid is transparent to electrons in the same way that glass is transparent to light: there are no processes by which the light can dribble away its energy in little pieces.

Imperfections in the crystal, represented by the substitution of atoms of a different kind for atoms of the pure metal, or by the restless thermal motion of the atoms around their "home" positions in the perfect-crystal lattice, scatter electrons and impede the flow of heat. Again the process can be compared to the incorporation of impurity atoms into an otherwise transparent glass: the impurities absorb light, and they color and even blacken the glass by means of this absorption. Thus the length of the path of an electron proves to be shorter in brass, in which zinc atoms have been added to copper, than it is in pure copper. The path length is also shorter at higher temperatures, because the random motion of the atoms at higher temperatures makes the substance more "turbid" to electron movement.

This discussion of the importance of the average length of the path traveled by the electron between collisions applies to the pushing of electrons by an electric field as well as to the pushing of electrons by a temperature difference. A good metallic conductor of electricity is therefore a good conductor of heat. Silver wins both races for the best conductor, with copper, gold and aluminum close behind. In fact, the thermal conductivity for all metals at the same temperature turns out to be equal to the electrical conductivity of the metal multiplied by a constant.

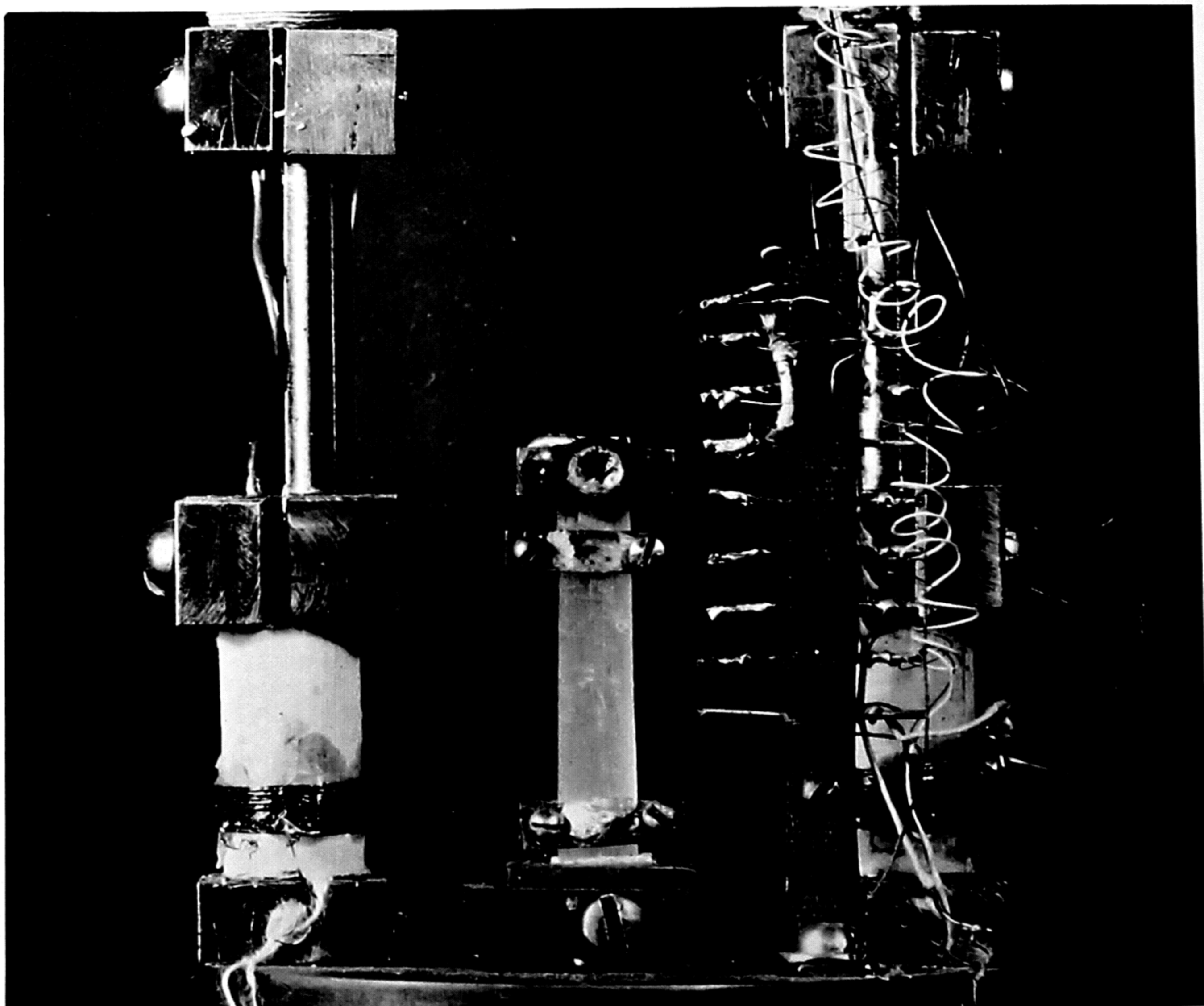
The two properties, however, do not vary equally with change in temperature. As the temperature rises, the electrical conductivity of a metal decreases almost reciprocally with the absolute temperature. This is true because the atomic vibrations become larger in proportion to the absolute temperature and the electron paths become correspondingly shorter. The thermal conductivity, on the other hand, executes a neat balancing act that leaves it independent of temperature. The average length of the electron paths decreases, to be sure, as the temperature increases, but the amount of heat carried by each electron is proportional to the temperature.

The thermal conductivity is proportional to the product of these two quantities and so remains independent of temperature.

Proceeding now from metals, with their abundant supply of mobile electrons, to semiconductors, with fewer electrons, and ultimately to insulators, with virtually none, one finds that electrical conduction and thermal conduction part company. The electrical conductivity drops as the mobile electron concentration does and reaches remarkably small values for good insulators such as quartz: silver is about 10^{24} times better as a conductor of electricity

than quartz is. The thermal conductivity, however, shows no such drop in value. As the electron concentration decreases, the thermal conductivity falls to perhaps a hundredth or a thousandth of the values characteristic of metals and then falls no more. There are, in fact, no good solid heat insulators. An effective heat insulator is not a solid at all but a porous, low-density substance composed mostly of dead air space.

Clearly some process other than conduction by electrons must take over the task of carrying the heat in nonmetals. This process is even more intricate and interesting than conduction by electron motions. It is the conduction of heat by



CRYOSTAT for measuring the thermal conductivity of a crystal (*vertical translucent bar*) at very low temperatures is shown at about twice its actual size. Heat flows into the crystal from the

tiny electrical heater attached to the top of the crystal; identical thermometers are clamped near each end. In use the apparatus is enclosed in an evacuated copper can immersed in liquid helium.

the vibrations of the atoms themselves. This process also occurs in metals, but it is covered up by the more effective electron process.

The mechanism of heat transport by atomic vibrations is basically simple. Atoms in a solid are packed tightly together. If an atom is set moving back and forth when part of a solid is heated, the atom nudges its neighbors, which in turn transmit this motion to their neighbors. Atomic kinetic energy is thereby carried from the hot parts of the solid to the cold. On the macroscopic scale this flow of kinetic energy shows up as a flow of heat. The mechanism is identical with the transport of sound waves in a solid, since they too are carried by the pushing of atoms by one another. The typical vibration frequencies of the atoms in thermal motion, however, far exceed those of audible sound. In heat conduction, frequencies of 10^{13} cycles per second are common. This is 35 octaves above middle C!

The system of atoms behaves like a system of masses and springs. Each mass is the analogue of a nucleus with its accompanying tightly bound electrons, and the springs are the analogues of the interatomic forces. Each mass in such a system has an equilibrium position that is equivalent to the home position of an atom, and it will vibrate around this position if it is initially displaced. If a mass in one part of the network is shaken, its motion is communicated to the other masses. This is analogous to the flow of heat from a hot to a cold region of a solid. After the external shaking has ceased all the masses will ultimately acquire the same kinetic energy of vibration, just as ultimately all the atoms in a solid arrive at the same temperature, which is a measure of their kinetic energy of vibration.

These atomic vibrations actually occur in tiny bunches called phonons. A phonon is simply a pulse of sound waves, comparable to the pulse of water waves from a stone dropped into water. Its name expresses its similarity to the photon, which is a pulse of light waves. The basic quantum nature of matter insists that energy occurs only in indivisible little lumps—quanta. Phonons and photons are respectively the quanta of sound waves and light waves. Phonons, then, are the carriers of heat in nonmetallic solids.

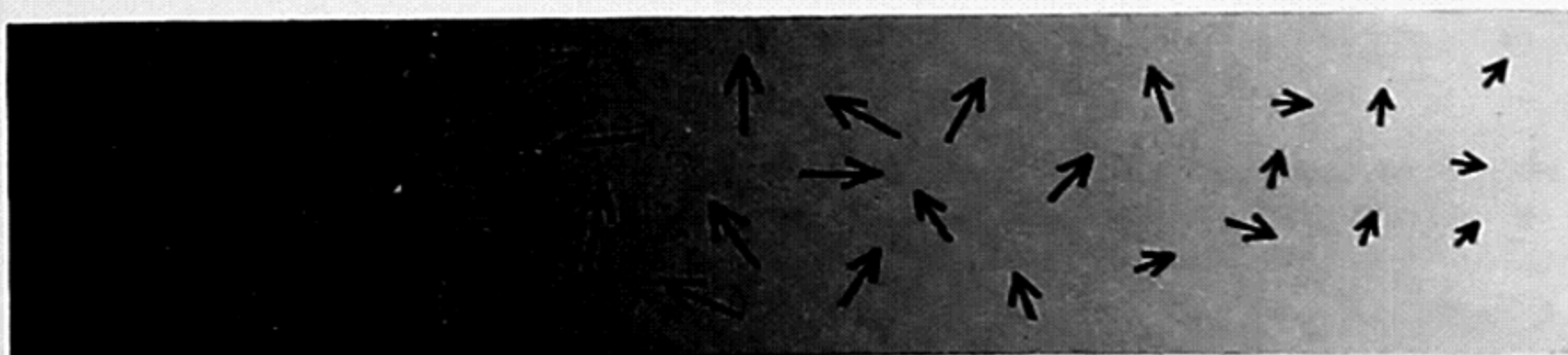
The thermal conductivity of a solid depends in a simple and almost obvious way on the properties of phonons. It is naturally proportional to the number of

phonons present, which increases rapidly with increasing temperature. It is proportional to the speed of the phonons, which is equal to the speed of sound waves in the solid and varies only slightly with temperature. Last and most important, the thermal conductivity is proportional to the free path that each phonon travels before it collides with some imperfection in the crystal. Such a collision, like the collisions of electrons with imperfections in the crystal lattice, reflects a phonon back toward the warm end of the solid. It is the variation of this free path from one temperature to another and from one solid to another that, more than any other factor, controls the thermal conductivity of all solids, metals as well as nonmetals.

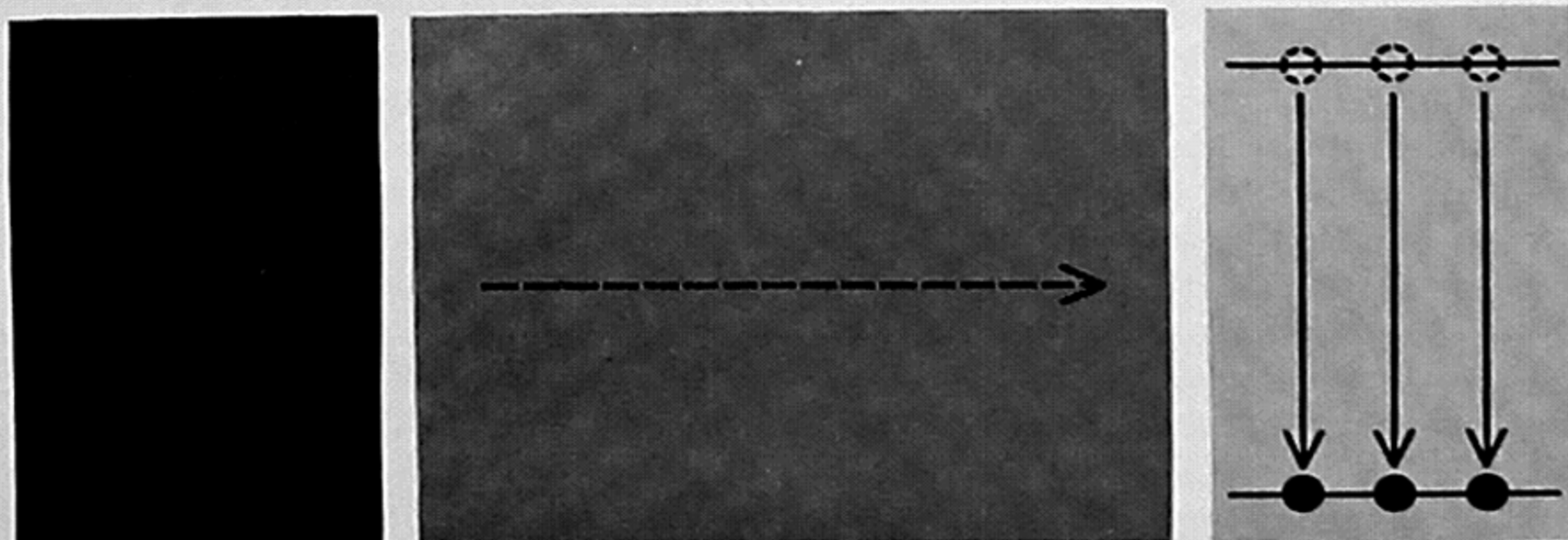
At ordinary temperatures solids are generously supplied with phonons. There are pulses of atomic vibrations moving in all directions. Paradoxically this abundance decreases the heat conductivity by phonons, which drops steadily as the temperature increases above room temperature. Although the number of phonons increases as the temperature rises, they impede one another's movements so seriously that their path lengths diminish faster than their population expands, in a manner suggesting the

overcrowding of a highway or of a cocktail party. When an atom is displaced from its home position by the passage of a phonon, another phonon encountering the misplaced atom is strongly diverted. The phonons themselves impede their own motions as they grow in numbers. At high temperatures phonons move scarcely farther than the distance between neighboring atoms before they are scattered.

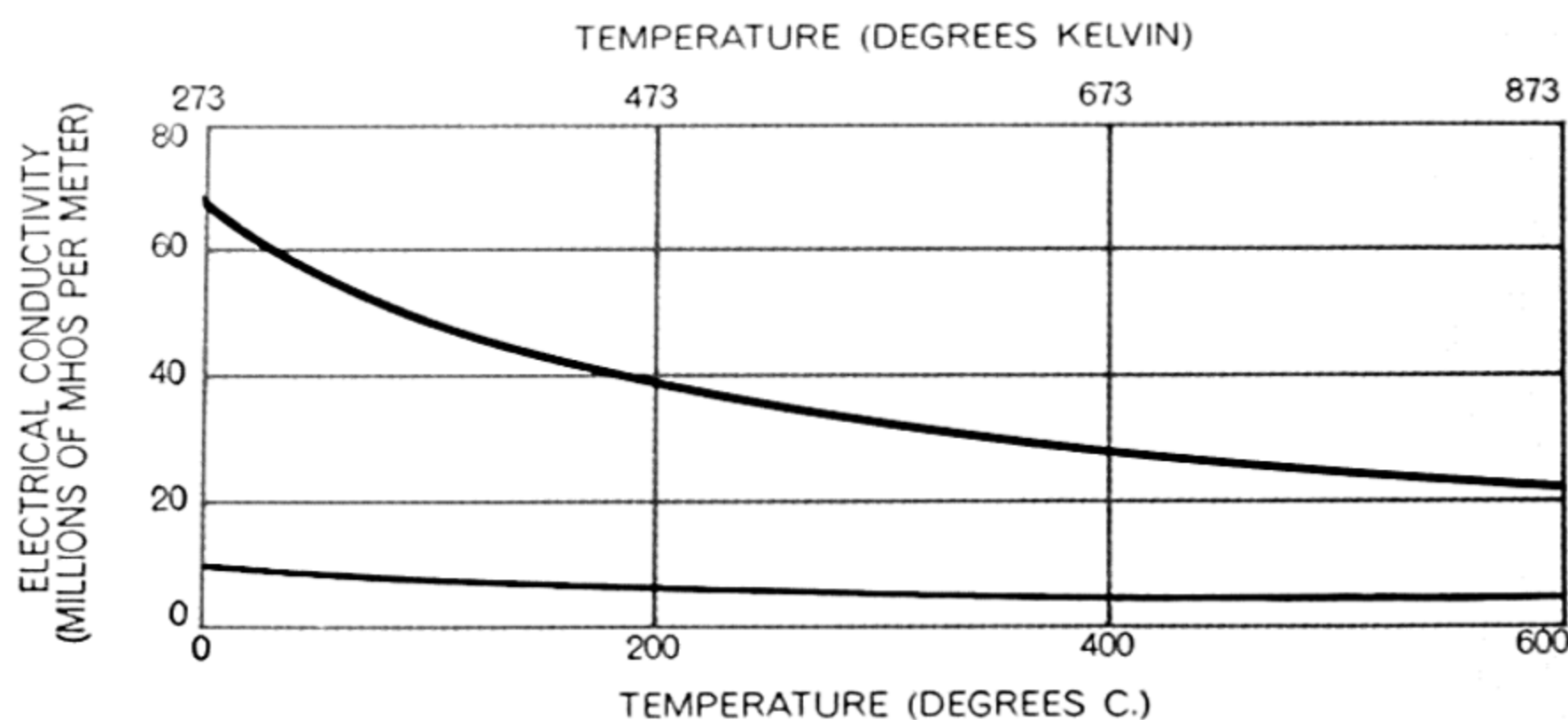
The presence of heavy atoms in a solid decreases the heat conductivity by lowering the velocity of sound and by enhancing the scattering of phonons. Designers of thermoelectric materials take advantage of this knowledge by employing heavy elements such as bismuth or tellurium, which have low heat conductivity, in compounds with metals that are rich in free electrons to provide high electrical conductivity. A major portion of the energy of heat, supplied to one end of such a thermoelectric generator, is carried by the free electrons to the other end in the form of electric current. The heavy atoms in the compound meanwhile impede the transfer of energy by the phonons, maintaining the temperature gradient between the hot and cold ends of the generator [see "The Revival of Thermoelectricity," by



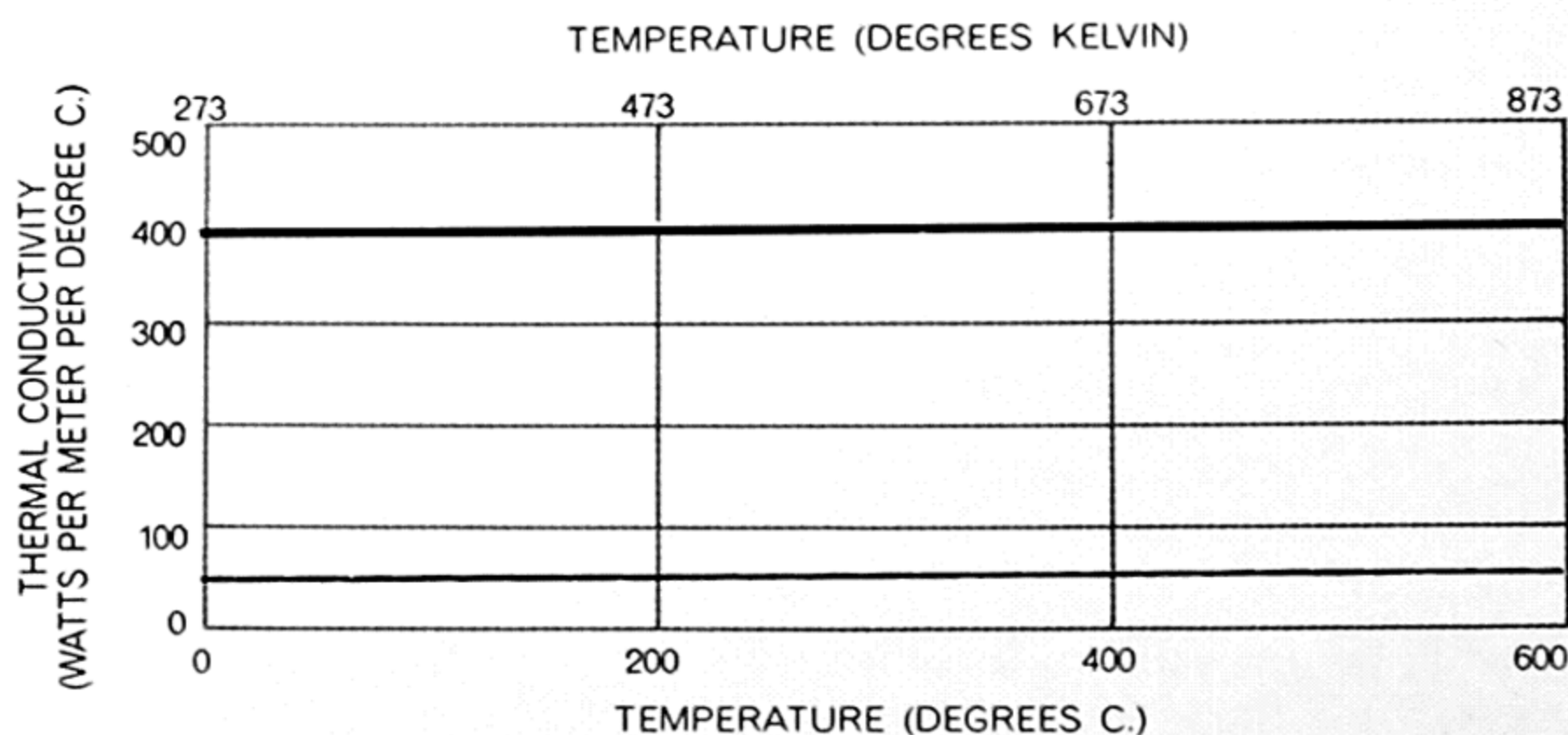
ELECTRONS IN METAL have greater speeds (i.e., greater kinetic energy) at higher temperatures. The flow of some of the more energetic electrons (represented by longer arrows) to the cooler part transfers their kinetic energy to that part, which is thereby heated.



THERMAL EXCITATION of electrons (black dots) in the hotter region of a solid raises them to higher energy levels and releases them from the atoms in which they were bound. Subsequent diffusion of these electrons to a cooler region is followed by the release of the excitation energy when they recombine with other atoms (i.e., fall to lower energy levels). This may be the dominant process of heat flow in nonmetals at extremely high temperatures.



ELECTRICAL CONDUCTIVITY in a metal falls as the temperature increases, because the mean free paths of the heat carriers (i.e., the electrons) are progressively reduced. The two metals represented by these electrical conductivity curves are silver (*black*) and iron (*gray*).



THERMAL CONDUCTIVITY of silver (*black*) and iron (*gray*), or of any metal, remains constant as the temperature rises, because the energy carried by each electron increases sufficiently to compensate for the reduction in the mean free paths of the electrons.

Abram F. Joffe; *SCIENTIFIC AMERICAN* Offprint 222].

It is at temperatures far below room temperature—temperatures near absolute zero, or minus 273 degrees centigrade—that the most striking phenomena of heat conduction occur. Although measurements in this regime date from the liquefaction of atmospheric gases and of helium half a century ago, only recently was the richness of the field of low-temperature studies of heat flow discovered by Robert Berman and others at the University of Oxford. At low temperatures solids such as rock-salt crystals exhibit heat conductivity that is as much as 400 times greater than it is at room temperature. Furthermore, the thermal conductivity at low temperatures is strikingly responsive to the addition of minute traces of impurities or other imperfections in crystals. Studies of ther-

mal conductivity in this temperature region thus powerfully assist the chemist, the physicist and the metallurgist in their investigations of the approach to perfection in crystals.

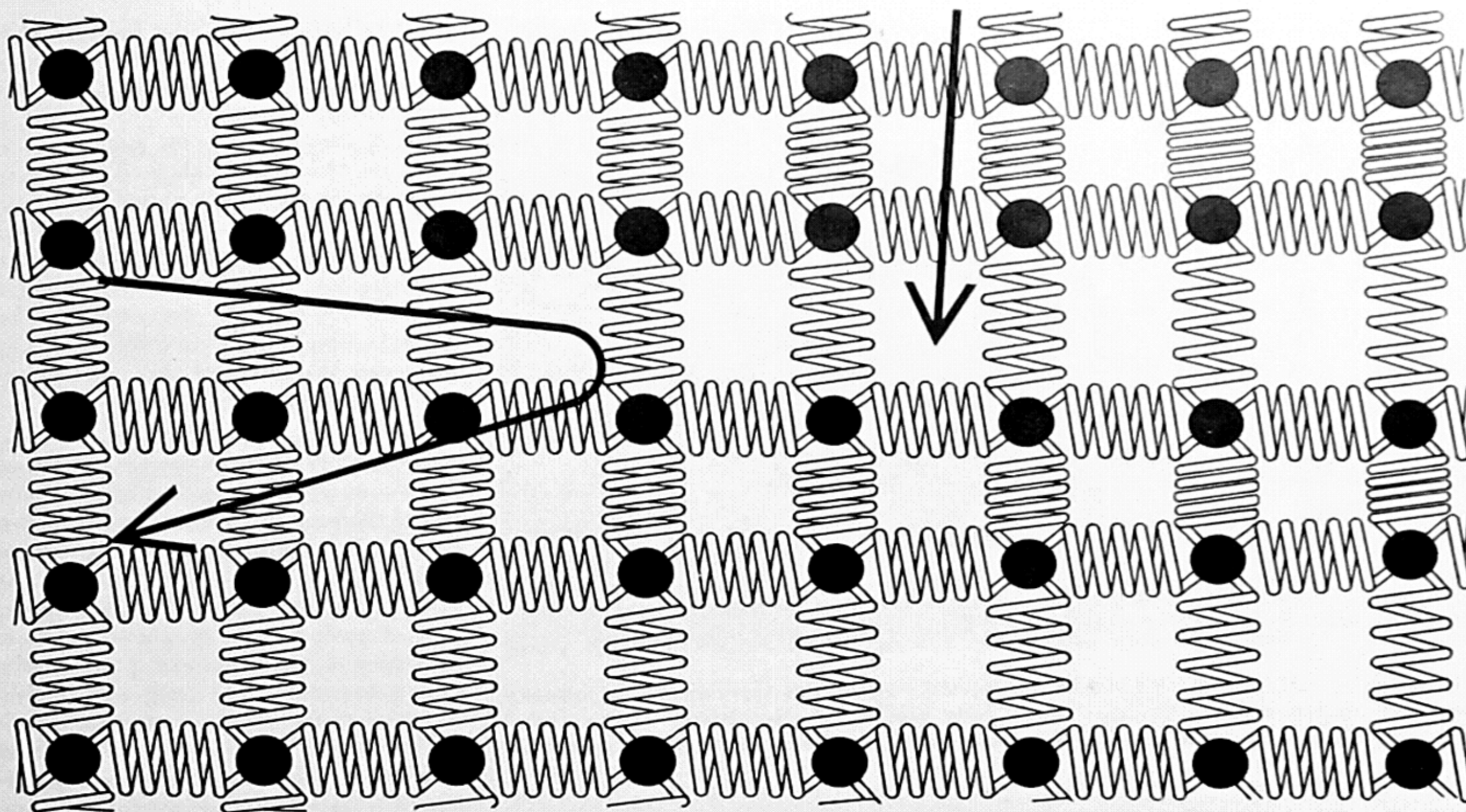
As the temperature drops, the thermal conductivity of nonmetals rises because the free path length of the phonons increases faster than the number of phonons decreases. Eventually a temperature is reached at which this path length—less than a millionth of an inch at room temperature—grows to lengths comparable to the dimensions of the specimen, typically an eighth or a quarter of an inch. Below this temperature the paths of the phonons are terminated by the edges of the crystal, and the phonons are reflected back into the interior of the crystal. The thermal conductivity drops sharply below this temperature; this is because the number of phonons decreases rapidly with further decline in

temperature, while the phonon path length and the phonon speed remain constant.

This interpretation of the increase and then the abrupt decrease in the conductivity of materials with decline in temperature was first derived in theory by H. B. G. Casimir of the University of Leiden from early measurements made by W. J. de Haas and Th. Biermasz of the same institution. A convincing demonstration of the validity of the proposed physical process has recently been given by Robert O. Pohl of Cornell University. He cut two specimens of different cross-sectional dimensions from the same single crystal of the salt lithium fluoride and found that the conductivity of the smaller specimen fell off sooner with the fall of temperature, in convincing agreement with Casimir's theory.

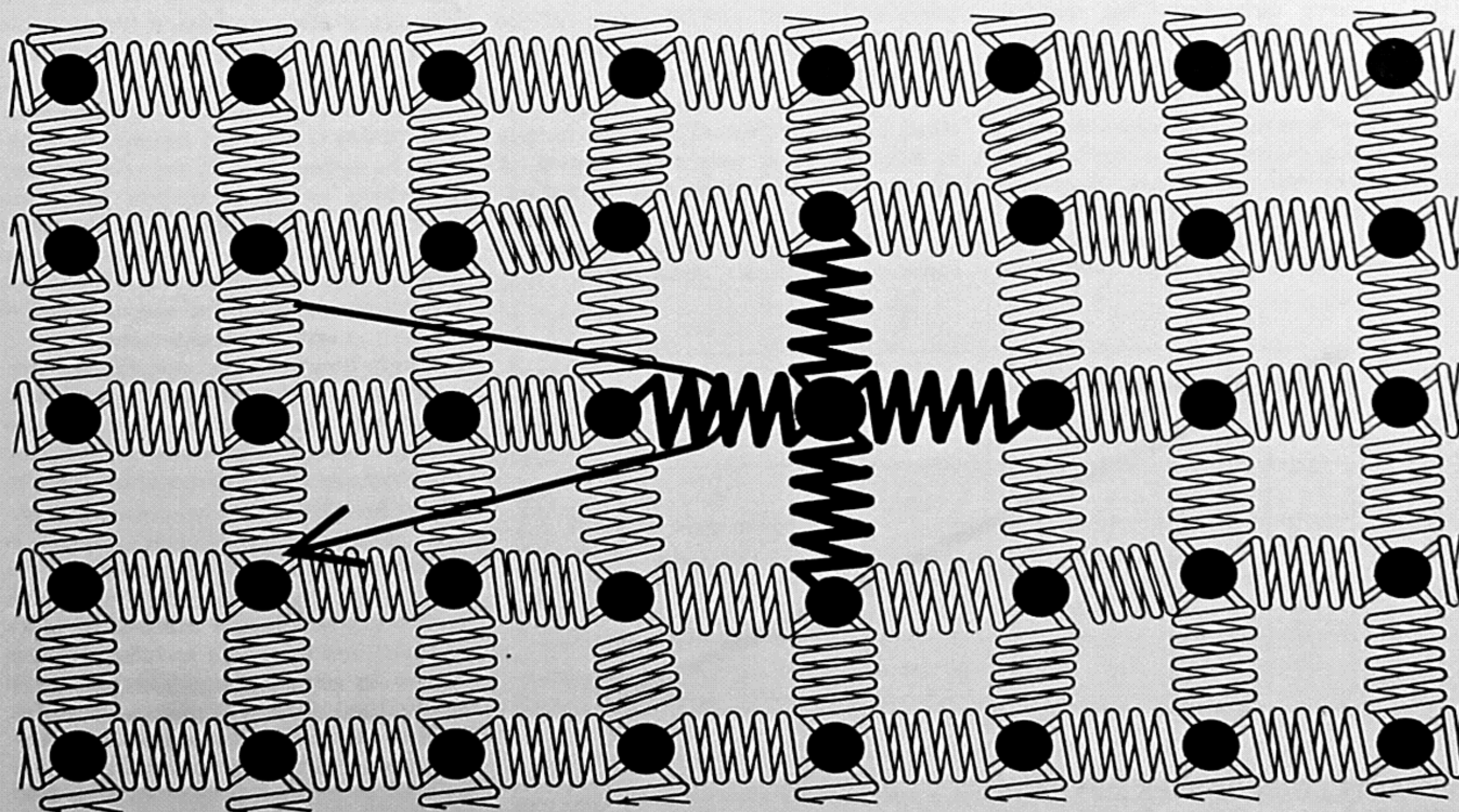
Recently Miles V. Klein, now at the University of Illinois but then at Cornell University, provided a dramatic demonstration of the fact that the thermal conductivity at low temperatures is highly sensitive to crystal imperfections. He measured the conductivity of crystals of sodium chloride (common salt) obtained from different sources in the form of large single crystals. Although each crystal was nominally pure and all were much purer than the "pure" chemicals on the shelf of a chemistry laboratory, the measured conductivities differed astonishingly. Since chemical impurity incorporated in a crystal scatters phonons, lower conductivity in certain of the specimens could be taken as evidence of imperfections in their crystal structures. Auxiliary chemical and physical experiments on these specimens indicated that an oxygen-containing radical, such as carbonate ion, was probably responsible for the depression of the conductivity in the less pure crystals. Klein intentionally "doped" crystals with oxygen-containing ions in known concentrations. One of these crystals, with only one part in 3,000 of this impurity, exhibited a conductivity comparable to that of the "pure" crystal with the lowest conductivity of the group. Therefore the purest of the crystals, with a low-temperature conductivity 100 times greater than that of the least pure, must have had less than one part in 300,000 of the impurity.

Under favorable conditions the thermal conductivity at low temperatures is sensitive to impurities even when they are present in concentrations of only one part per million. Since the atoms of



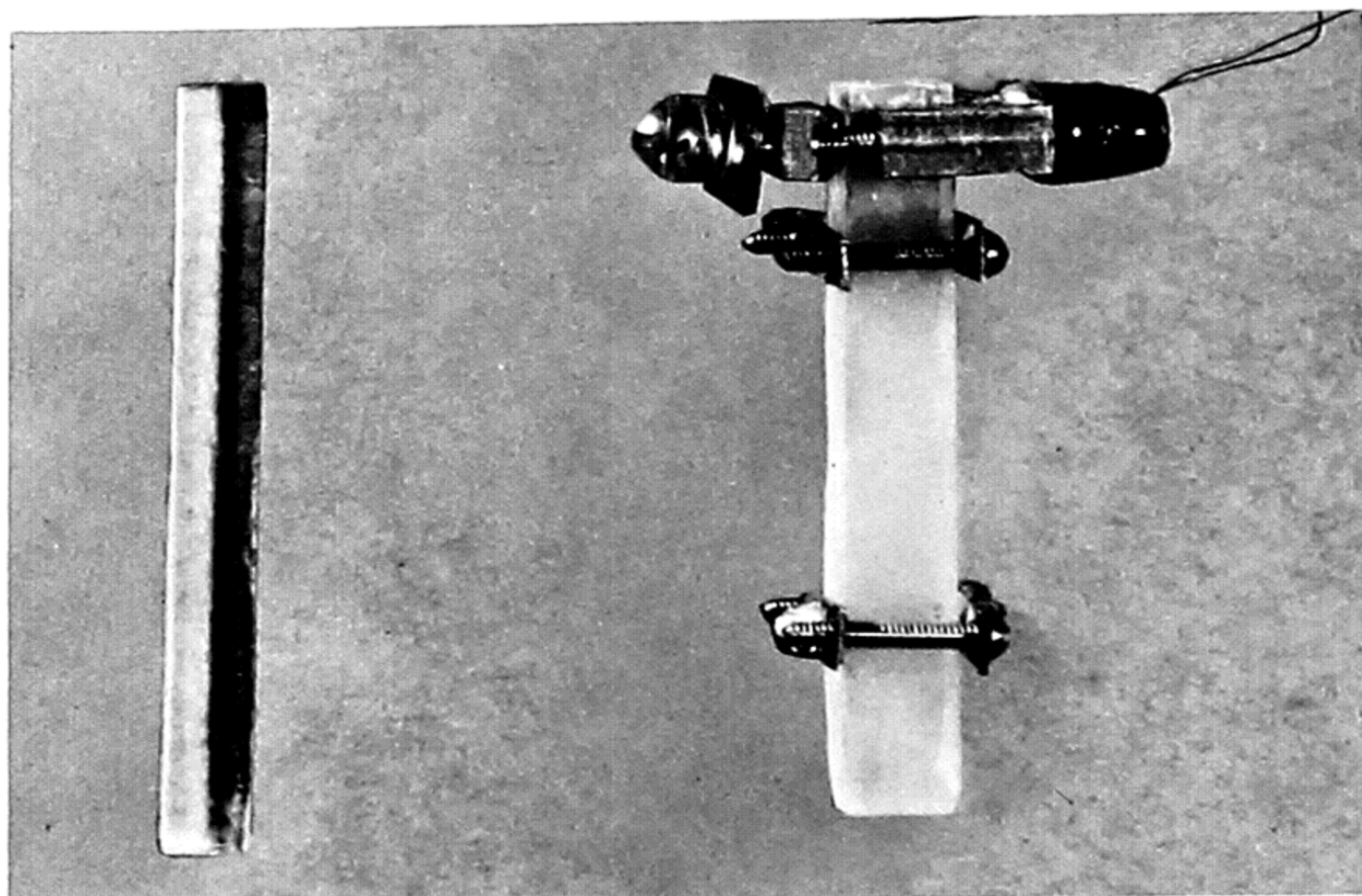
PHONON (*arrow at right*) displaces atoms from their normal positions. In the mechanical model of atoms in a solid depicted here, the atoms are represented by masses (*gray circles*) and the interatomic forces by springs. Another phonon arriving at this time

would encounter a region of imperfection; it would be strongly reflected (*curved arrow at left*), its free path terminated and the flow of heat impeded. Such phonon-phonon scattering limits the thermal conductivity of materials at ordinary and high temperatures.



IMPURITY IN CRYSTAL LATTICE introduces a different mass (*large gray circle*) and a different interatomic force (*gray springs*). An incident phonon is strongly reflected (*curved arrow*). Phonon-

imperfection scattering is particularly prominent in limiting thermal conductivity at low temperatures; with large impurity concentrations it can be an important factor at room temperatures.



POTASSIUM CHLORIDE CRYSTALS for thermal conductivity experiments are approximately 40 millimeters long. The specimen at right, with an electrical heater (a wire-wound resistor) and two clamps attached to it, is ready to be mounted in the cryostat apparatus.

an element in a crystal can be those of two or more isotopes with slightly different masses, the purest crystal may be composed of atoms of varying masses. Glen A. Slack, now at the General Electric Research Laboratory but then at Cornell, first discovered that even this small variation produced an observable depression of the thermal conductivity.

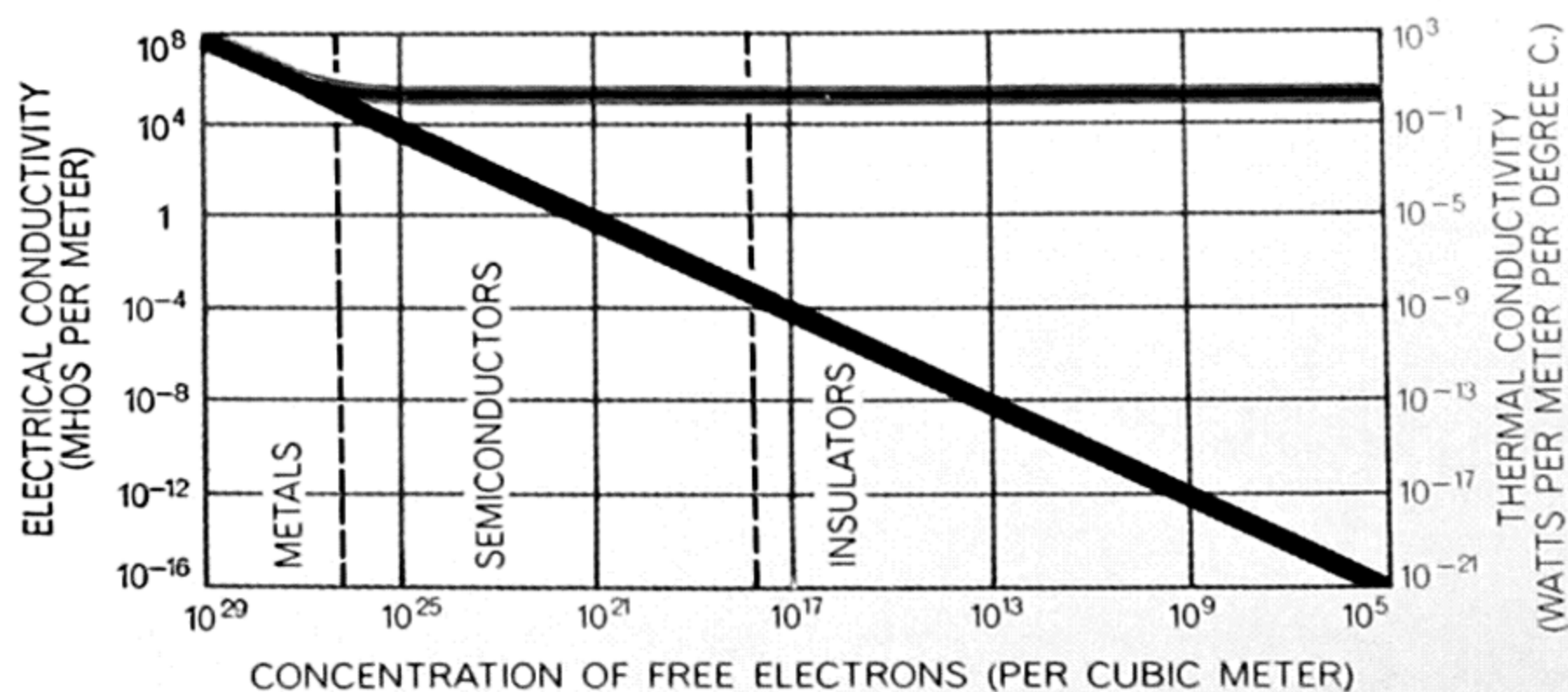
Thermal conductivity measurements are not only sensitive as a method for studying crystal perfection; they are also selective. When different kinds of imperfections limit the phonon paths in a crystal, the thermal conductivity be-

haves differently as a function of temperature. Pohl has irradiated a lithium fluoride crystal with X rays, thereby creating imperfections known as F-centers, in a concentration of about one F-center for each 100,000 atoms of the crystal. An F-center is a site that would be occupied by a fluorine ion if the crystal were perfect but that is occupied instead by an electron. This imperfection is localized in a very small region of the crystal and approximates a point imperfection. Another group at Cornell has squeezed lithium fluoride crystals in order to introduce "dislocations," that

is, imperfections resulting from the slippage of whole rows and planes of atoms during deformation. Comparison of the variations in the thermal conductivity of crystals flawed by point imperfections and by dislocations showed that conductivity in each case varied with temperature in a characteristic way. Thus in addition to the mere detection of an imperfection, it is frequently possible to learn much about the nature of the imperfection in an "unknown" crystal by analyzing the curve of thermal conductivity as a function of temperature.

In all of these experiments there are two considerations that compel the use of the low-temperature regime, even though it requires complex apparatus and the handling of liquid helium. At low temperatures, first of all, the limitation of path length by phonon-phonon crowding is not important; path lengths hence become quite long and highly sensitive to the presence of minute concentrations of imperfections. Second, in order to explore temperature dependence it is necessary to cover a wide range of temperature, say a factor of 100 in absolute temperature. Such a range is available only by "dividing" room temperature, that is, by going downward on the temperature scale to within a few degrees of absolute zero; it is impossible to multiply room temperature 10 times or so on the absolute scale (from 300 degrees centigrade above absolute zero, say, to 3,000 degrees) before approaching the upper limit of the temperatures at which materials remain solid. Information gained about phonon-transport processes from low-temperature experiments has already advanced the design of materials for use in practical devices at ordinary temperatures.

High temperatures nonetheless present intriguing problems in heat flow in nonmetals. Since thermal conduction by phonon transport becomes steadily less effective as the temperature climbs, other heat-transport processes eventually dominate. One such process is ordinary radiant-heat transport: the mechanism by which the sun's heat comes to earth across millions of miles of space and by which an infrared lamp warms its surroundings. This process is simply the transport of photons of visible light or of infrared radiation. The transport proceeds through transparent solids as well as through a vacuum. The heat radiated through a transparent quartz rod when one end is heated to a high temperature vastly exceeds the heat carried by phonons. As glass blowers



CONCENTRATION OF FREE ELECTRONS in a solid directly affects its electrical conductivity; that is, as this concentration decreases (proceeding from metals through semiconductors to electrical insulators), the electrical conductivity (black curve) decreases correspondingly. Thermal conductivity (colored curve) is similarly affected at first, but heat transport by atomic vibrations soon begins to dominate, preventing any further decrease.

are well aware, the radiant heat is "piped" by internal reflection within the rod; although the walls of the rod may feel only warm, a bad burn results if the radiation emerging from the unheated end falls on one's hand. At sufficiently high temperatures this radiation process becomes important even in an opaque material such as ordinary china. In such a translucent material the free paths of the photons are short but not infinitesimal, and heat is still transferred by radiation.

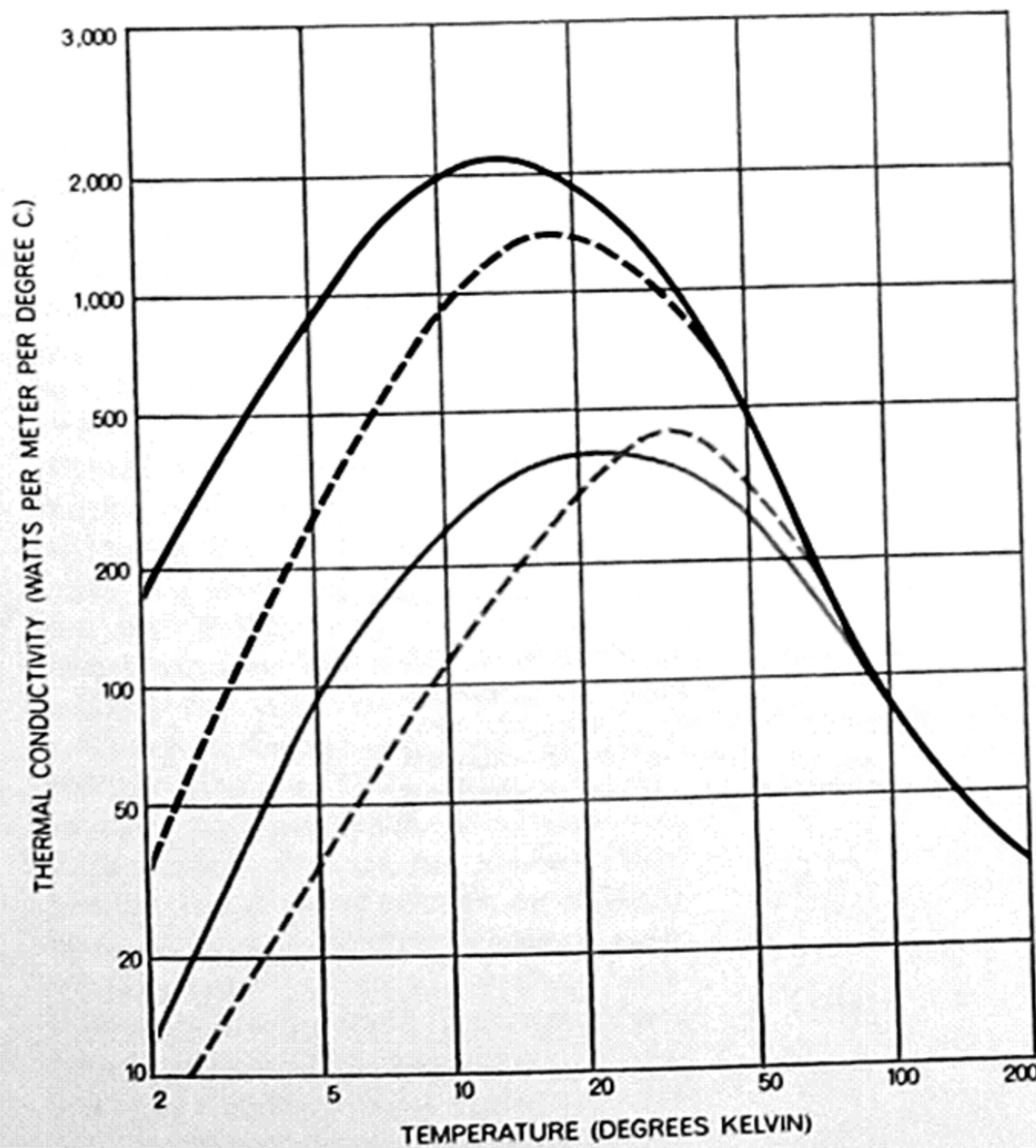
Another heat-conducting process arises at high temperatures in nonmetals from an electron energy-transport mechanism that is different from free-electron flow. Nonmetals do not have anything like the concentrations of free electrons (several per atom) that metals have. A few electrons, however, can be removed from

atoms if the temperature goes high enough. The energy required to pull an electron away from an atom is many times greater than the kinetic energy of motion of a typical free electron. Electrons thermally released from atoms at the hot end of a bar of nonmetallic solid can diffuse toward the cooler end, just as in a metal. At this point they can again combine with atoms, releasing the same amount of energy that was supplied to them at the hot end. Although these electrons have carried some kinetic energy, as in a metal, the far more important burden they have carried is the energy required to release them from atoms. Hence at high temperatures a relatively few electrons in a nonmetal, each carrying this large energy, can be as effective heat carriers as the electrons in a metal.

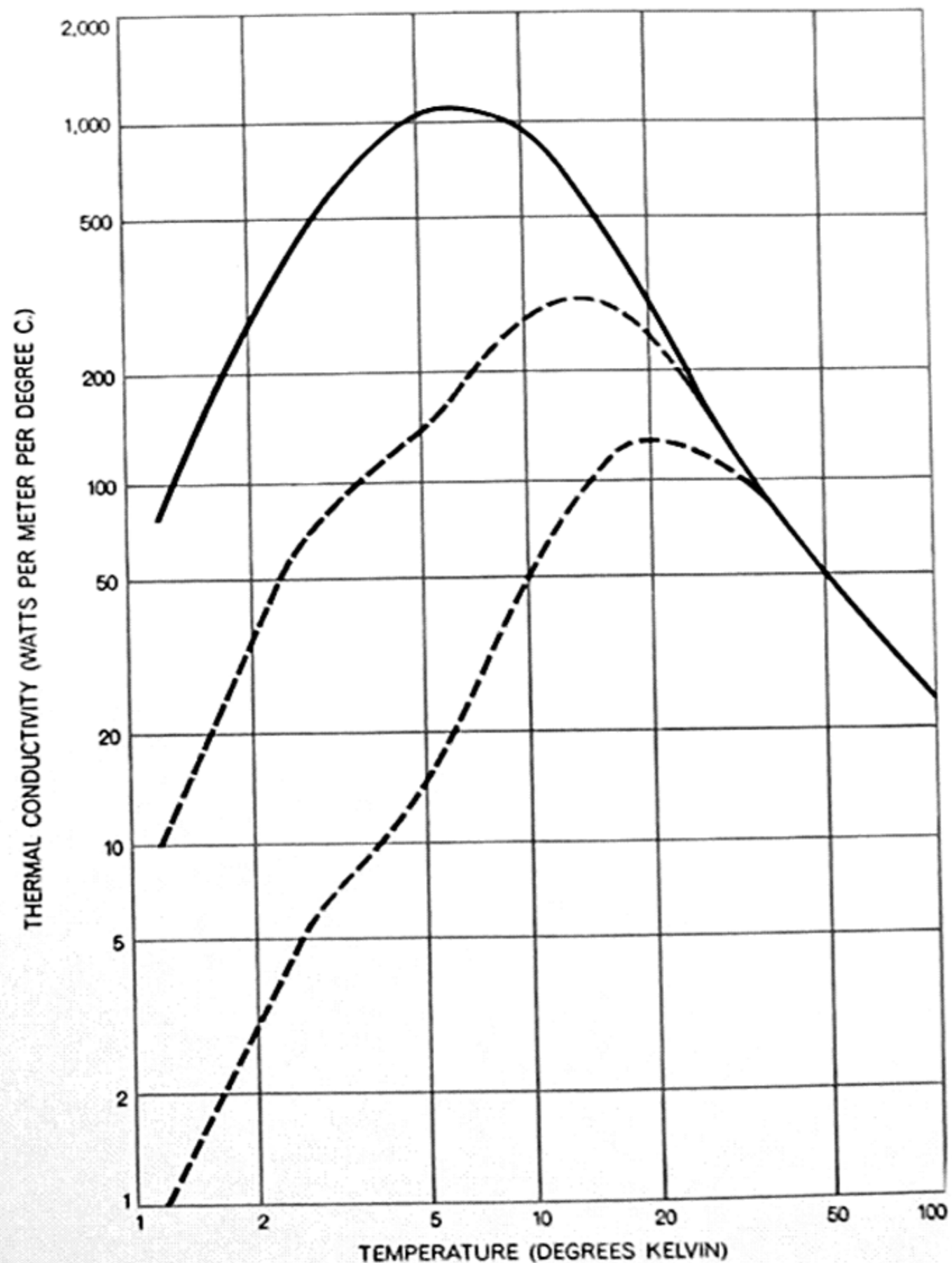
It is not easy to determine whether

radiation or excitation transport is the more important in a given material at a high temperature. Measurements are difficult, largely because of errors induced by radiant heat gain or loss at the surface of the test specimen. Yet more measurements and understanding are urgently needed to design such devices as thermoelectric power generators.

Thus the venerable old experiment of measuring the thermal conductivity of a solid has acquired new vitality. Part of the renewed vigor springs from the need for exotic materials with controlled conductivities at extremes of temperature. But a major part arises from the power and versatility of low-temperature thermal conductivity measurements for the study of physical processes in solids.



SIZE AND IMPERFECTIONS affect thermal conductivity differently. The only difference between the crystals represented by the two curves at top is that one (*solid black line*) is larger than the other (*broken black line*). Below a temperature of about 50 degrees Kelvin the paths of the phonons are terminated by collisions with the crystal surfaces; as a result the smaller crystal has the lower conductivity. The chief difference between the larger crystal and those represented by the two curves at bottom is that the latter contain either numerous F-centers (*solid gray line*) or dislocations (*broken gray line*), whereas the former has had no imperfections added.



THREE SODIUM CHLORIDE CRYSTALS, all chemically pure, nonetheless show considerable differences in their thermal conductivity. The impurities present in two commercially available crystals (*two bottom curves*) depress their thermal conductivity by as much as a factor of 100. The third crystal (*solid black curve*) is a specimen grown for experimental purposes from highly purified salt.

The Author

ROBERT L. SPROULL directs the Materials Science Center at Cornell University, where he is also professor of physics. After obtaining his B.A. at Cornell in 1940 and a Ph.D. in physics three years later, Sproull did research on microwave electronics from 1943 to 1946 as a research physicist with the Radio Corporation of America; he joined the Cornell faculty in 1946. Sproull spent 1952 at the Oak Ridge National Laboratory, served as editor of the *Journal of Applied Physics* from 1954 to 1957, and in 1958 and 1959 he was physicist with the Brussels firm European Research Associates. He became director of the Materials Science Center in 1960.

Bibliography

THERMAL CONDUCTION IN SEMICONDUCTORS. J. R. Drabble and H. J. Goldsmid. Pergamon Books. The Macmillan Company, 1961.

THERMAL CONDUCTIVITY. Roger J. Runck in *High Temperature Technology*, edited by I. E. Campbell. John Wiley & Sons, Inc., 1956.

THERMAL CONDUCTIVITY IN SOLIDS. J. A. Krumhansl and W. S. Williams in *Thermoelectricity*, edited by P. H. Egli. John Wiley & Sons, Inc., 1960.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

THE PHYSICS OF VIOLINS

by Carleen Maley Hutchins

Modern acoustics is making it possible to account for the exquisite performance of the violins made by the Italian masters. The results promise a further evolution in the instruments of the violin family.

During the Renaissance there grew up in Italy two new families of musical instruments, both of them stemming from primitive stringed instruments of the Middle Ages such as the rebec and lute. The earlier of the two groups to emerge was the viols; the later, following about a century afterward, was the violins. The violin was not an outgrowth of the viol but a somewhat later development from similar sources, and the two were lively competitors for a long time. Composers wrote distinctive music for each kind of instrument, and each had its virtuoso performers. Eventually the violin family, having a richer and more powerful sound, supplanted the older group—except for the largest and lowest-pitched instrument, which survives as the bass viol.

As this story unfolds it will be clear that viols still have more than mere historical interest. For the moment I shall describe the viols briefly. They came in a variety of shapes and sizes [see illustration on page 772], most of them having a flat back unlike the beautifully arched back plate of the violins. They had five, six or more strings, more slackly tuned than violin strings and supported on a flatter bridge. Often their finger boards were crossed by gut frets resembling the metal ridges on the finger board of a guitar. Their wooden sounding boards were lighter and more flexible than those of the violin family.

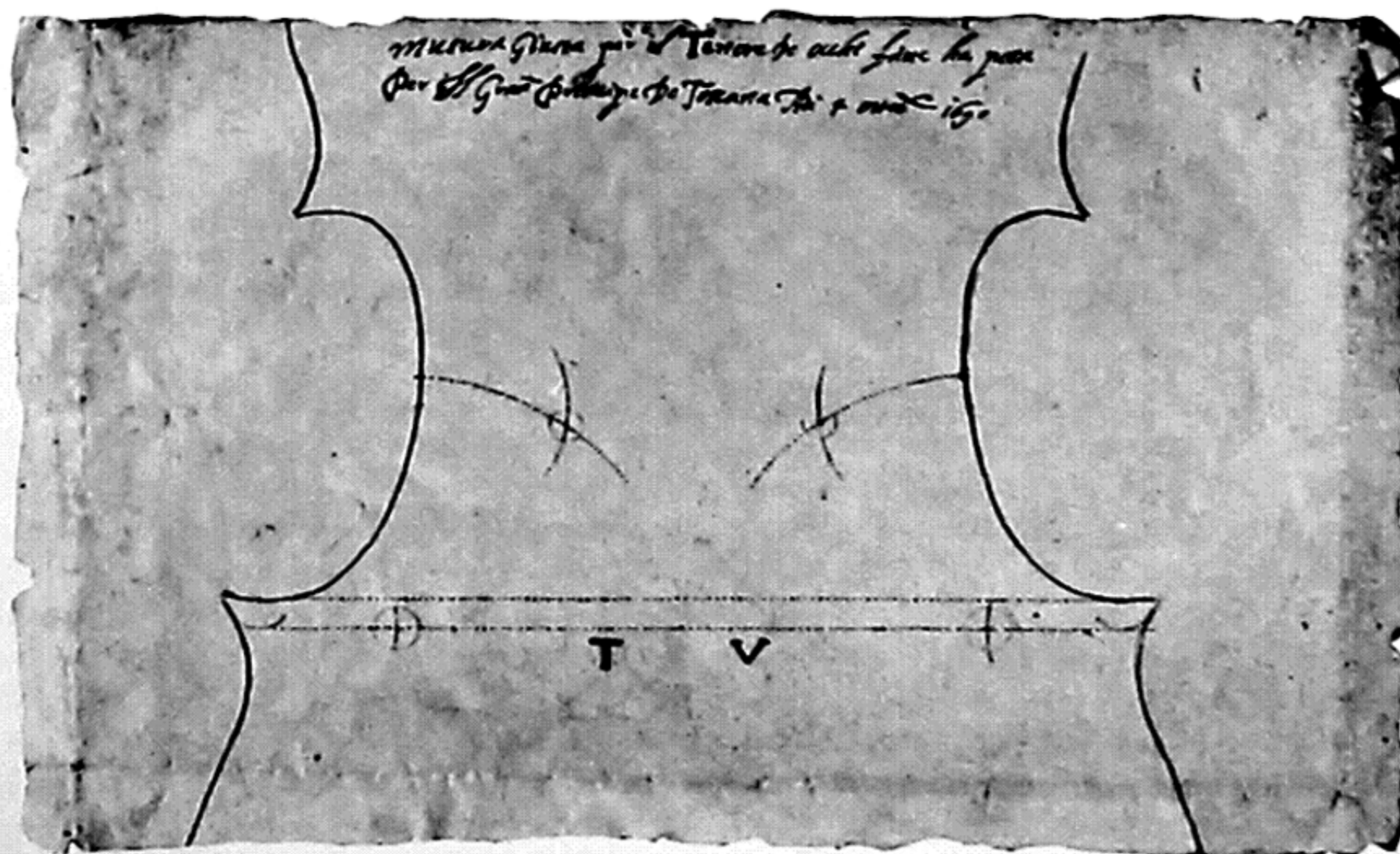
Exactly who invented the violin is not clear. It may have been Andrea Amati, who in any case founded the great Cremona school of violinmakers. Amati died around 1580; within 150 years or so his descendants and their pupils, particularly Antonio Stradivari and Giuseppe Guarneri, had brought the art of violinmaking to such an extraordinarily high level that it is only now that one dares

to dream of equaling or surpassing it. These early masters must have had an open mind toward the little that was known in their time about the physics of sound. Their successors deserve credit for having lovingly preserved an art, but certainly not for advancing a science. In effect they have formed a cult that has been plagued with more peculiar notions and pseudo science than even medicine.

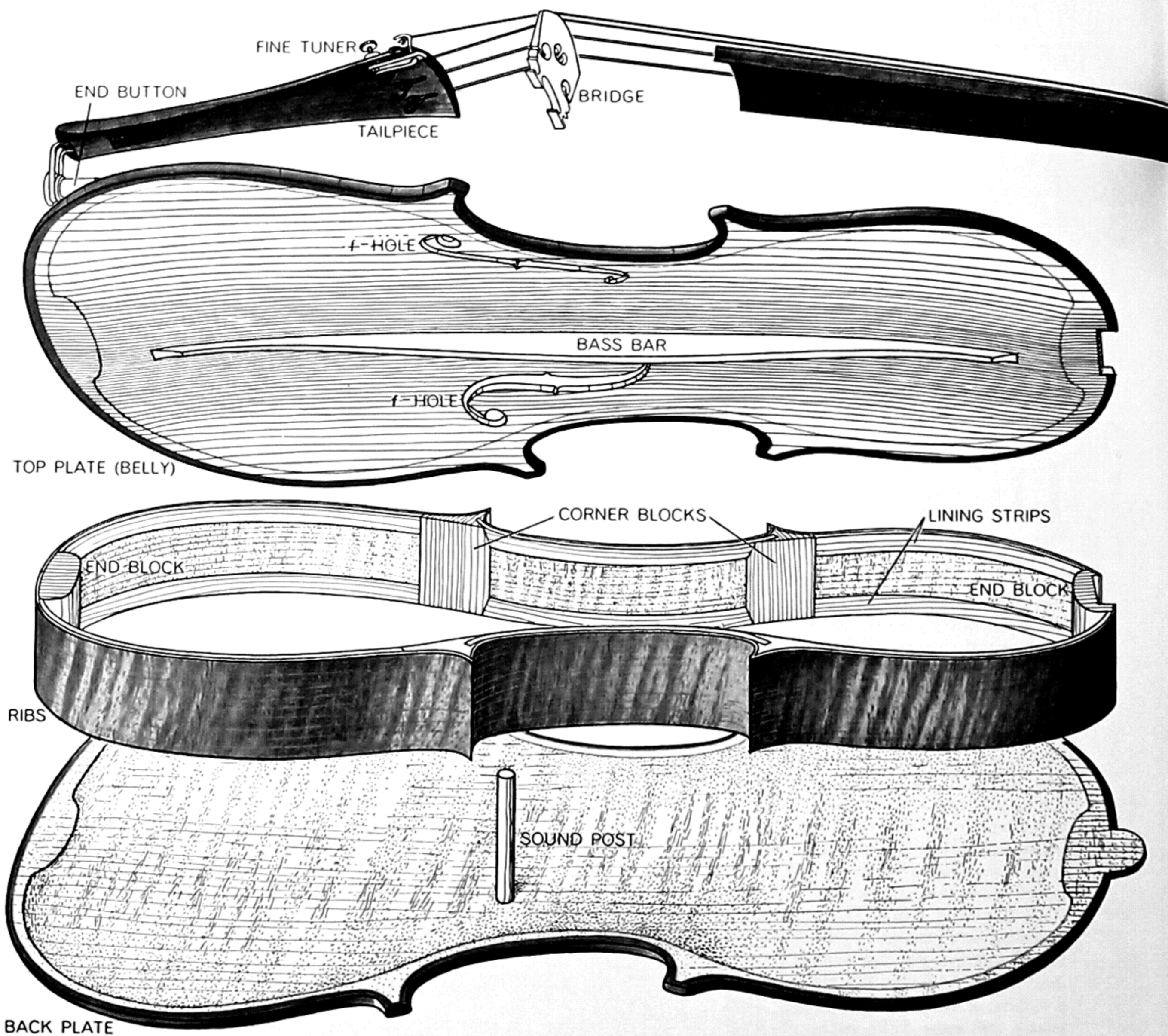
Today the well-developed science of acoustics is applicable to the understanding and making of violins. For the past 30 years or so a handful of interested physicists, chemists, musicians and some people who, like me, began as amateurs have been applying it. In fact, we have organized ourselves informally as the Catgut Acoustical Society. Much of what

has been learned is still empirical, but it is nonetheless interesting and valuable. In this article I shall try to touch on at least the high spots of our studies.

In essence a violin—as well as its larger, deeper-voiced relatives, the viola and the cello (properly the violoncello)—is a set of strings mounted on a wooden box containing an almost closed air space. Some energy from the vibrations induced by drawing a bow across the strings (precious little energy, it turns out) is communicated to the box and the air space, in which are set up corresponding vibrations. These in turn set the air between the instrument and the listener into vibration; in other words, they produce the sound waves that reach his ears. That is the main story. The sound of a violin, putting aside the acoustics of



DRAWING BY ANTONIO STRADIVARI marks positions of the upper and lower ends of "f-holes" in a tenor viola. At top he has written: "Exact measurements for the sound holes of the tenor made expressly for the Grand Prince of Tuscany, the 4th day of October, 1690."



ANATOMY OF VIOLIN INSTRUMENTS is essentially the same for the violin, viola and cello. The exploded view of the viola on these two pages shows the top plate, ribs, back plate and devices

for stringing at left, the neck, scroll and finger board at top right. Immediately below, a section of the top plate illustrates the bilateral symmetry of the grain of the spruce wood. At bottom

the room in which it is played and the skill of the player, depends on the transfer of vibration from string to sounding box to air.

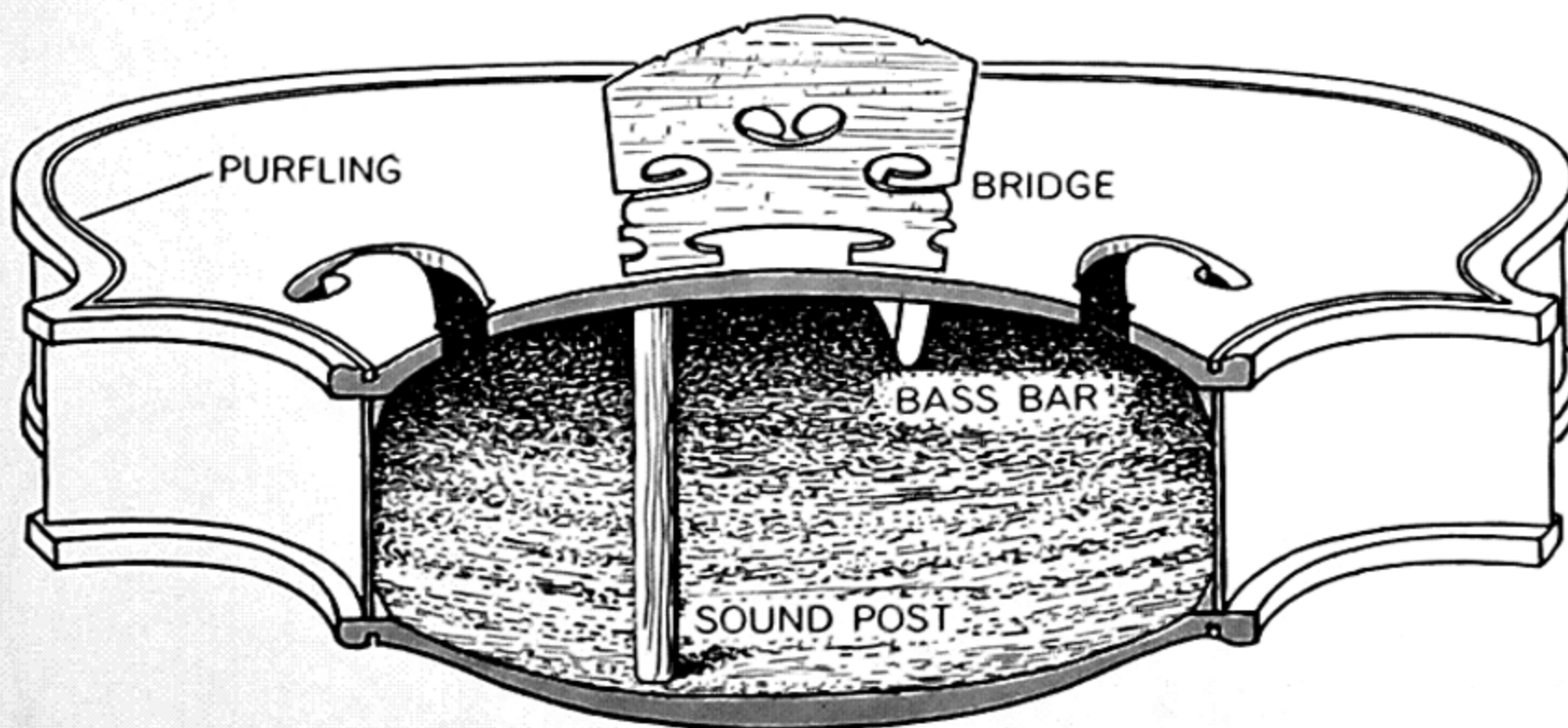
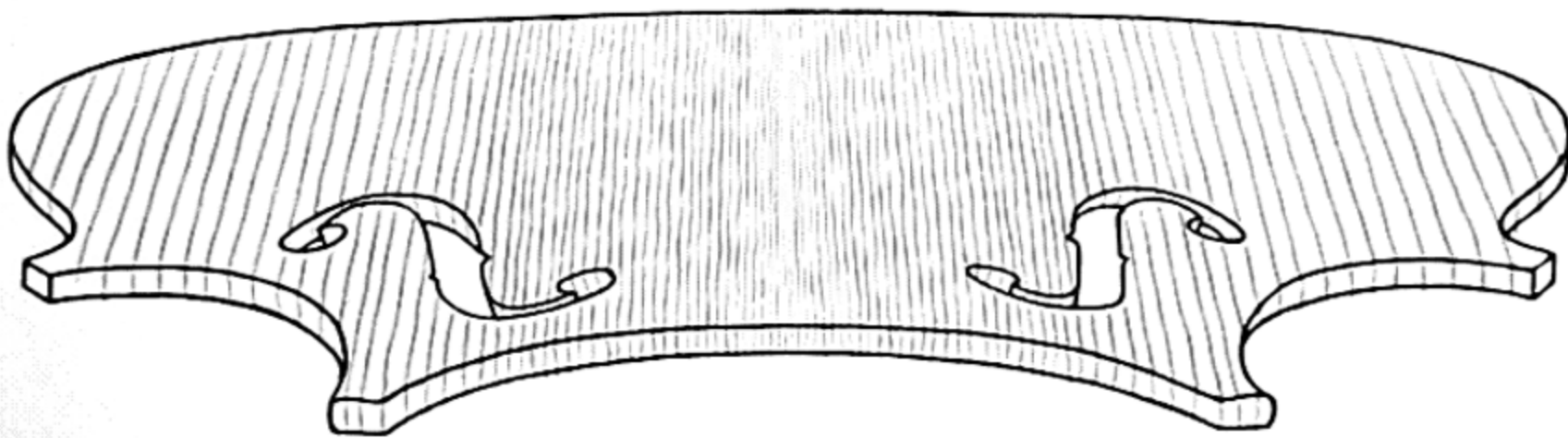
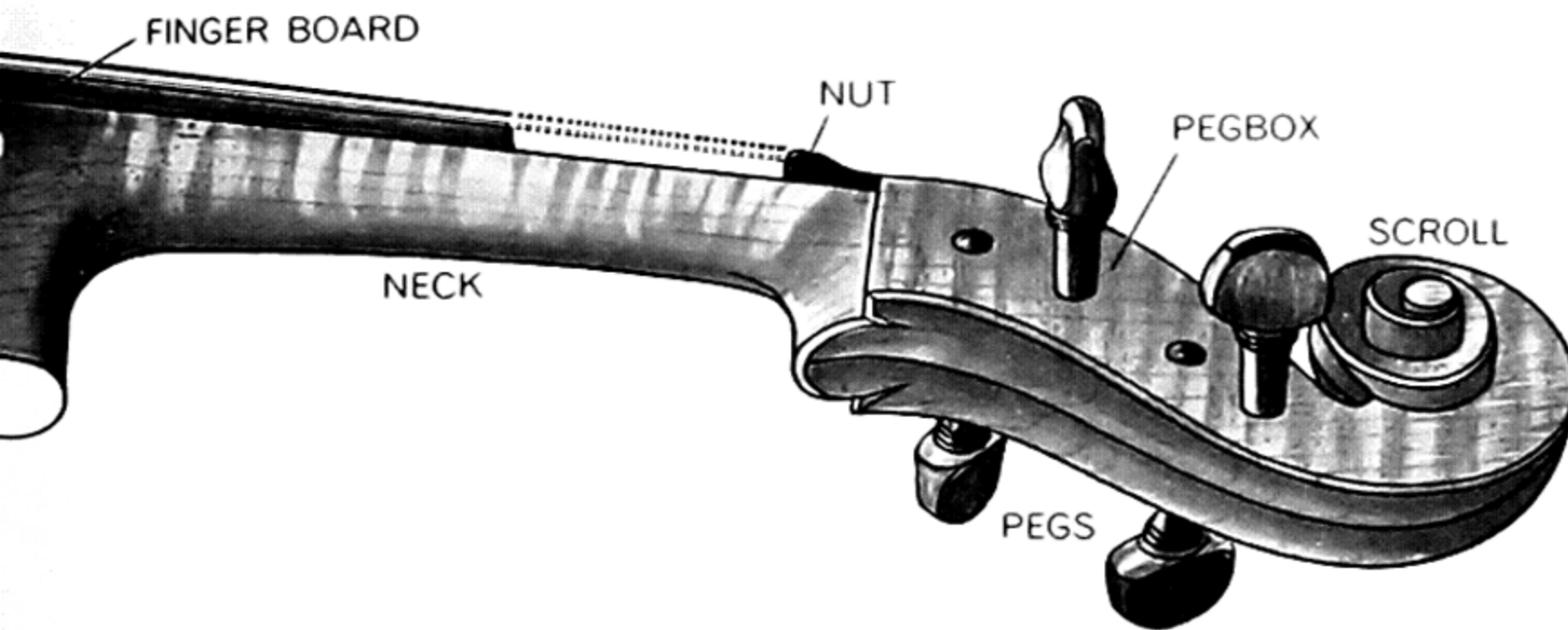
The Basic Violin

Before getting into this apparently innocent problem, which turns out to be a veritable jungle of unknowns, it is worthwhile to pause for a moment to examine the instrument itself. Violin strings are usually made of metal, pig gut or gut wound with fine silver or alumi-

num wire. The sounding box consists of a front plate and a back plate, both arched slightly outward to form broad bell-like shapes, and the supporting ribs, or sides. The back plate is carved with chisel, plane and scraper, traditionally from a block of curly maple seasoned for at least 10 years and not kiln-dried. (Pear or sycamore wood are sometimes used.) It can be a single piece or two pieces carefully joined. In thickness the back plate varies from about six millimeters in the center to almost two millimeters just inside the edges (from

1/4 inch to 5/64 inch). The sides are pieces of matching curly maple, thinned down to a millimeter all over, bent into shape and glued to spruce or willow blocks set in the corners and at the forward and rear ends of the plates.

The top plate, usually spruce, is split lengthwise from a log and then joined so that the wood of the outside of the tree is in the center of the top, making the grain bilaterally symmetrical. In thickness the top plate ranges from two to three millimeters, and a pair of beautifully shaped "f-holes" are cut into each



right a cross section through the middle of the instrument shows the relative positions of the bridge, bass bar and sound post. The purfling consists of three very narrow strips of wood that are set in a shallow groove around the edge of both the top and back plates.

side of the plate. All around the outside of each plate, near the edge, is cut a shallow groove in which is inlaid the "purfling," consisting usually of two strips of black-dyed pearwood and a strip of white poplar.

Other materials may be mentioned: curly maple for the neck, ebony for the finger board, rosewood or ebony for the tuning pegs and tailpiece, hard maple for the bridge. The outside of the instrument is treated with filler and varnished. Filler, varnish and glue all contribute to the over-all characteristics of the violin,

but there is no definite evidence to show that 300 years ago any of them was superior to the materials available now. In fact, the Catgut Acoustical Society is working to discover new substances that may be even more effective than the old. But it is a slow, painstaking search.

These were the general specifications for the Cremona fiddles and, with minor variations, for all good instruments since. Whether there is a mysteriously unique virtue in any of the woods or finishes, or whether some other types might not do as well for various purposes, is an

open question. In a few years we hope to have some answers.

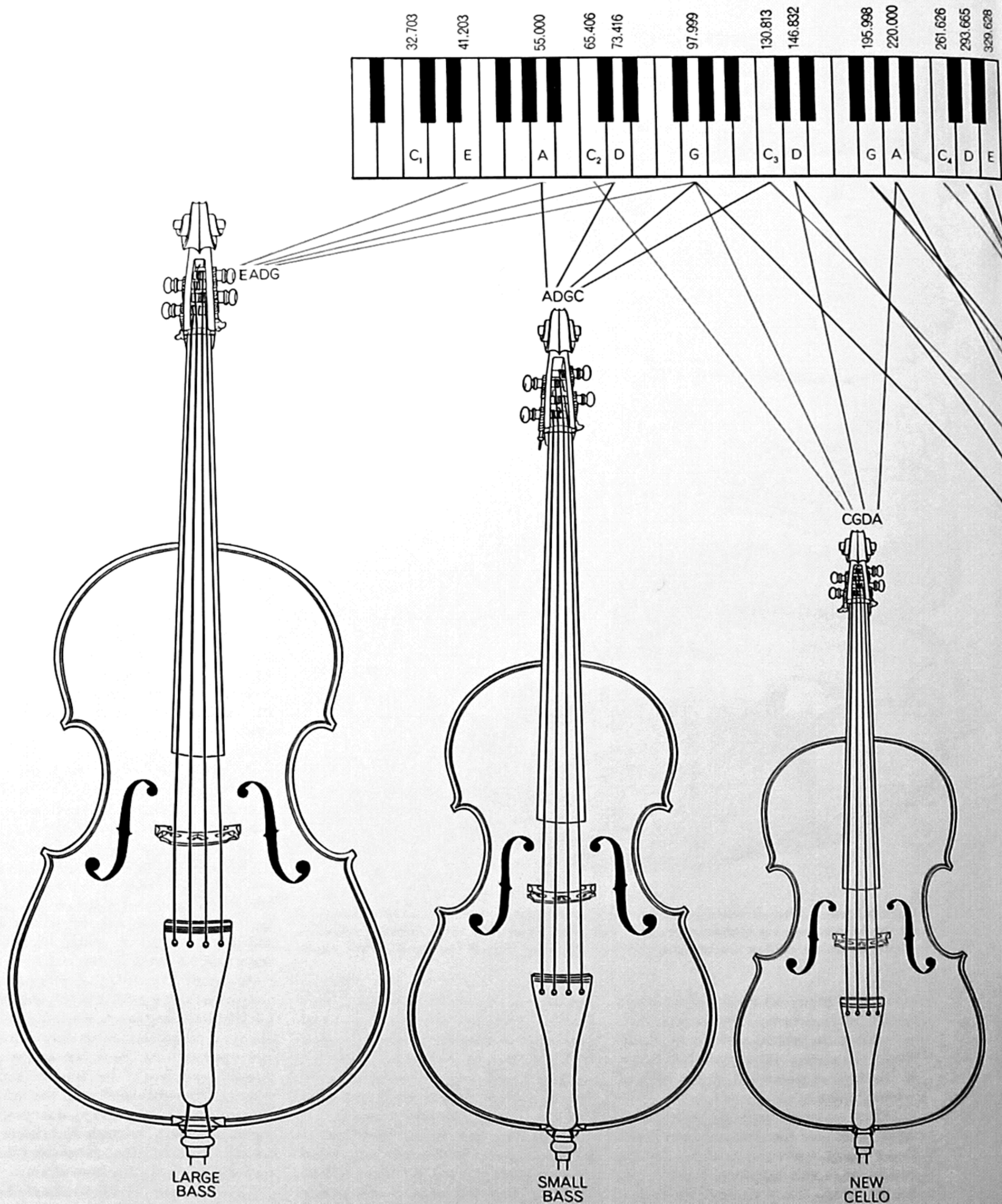
I might anticipate a bit here to mention a point that illustrates the subtlety of some of the problems in understanding the violin. Does the purfling serve any purpose other than decoration? It happens that the wood of the plates underneath the purfling is extremely thin. After years of playing, the glue that holds the purfling strips in their grooves begins to crack, in effect creating a vibrating plate with very thin edges. Frederick A. Saunders, professor emeritus of physics at Harvard University, has suggested that this may be a factor in the improved tone of an instrument that has been played for a long time.

The combined tension of the four strings of a properly tuned violin comes to around 50 pounds. As a result about 20 pounds is directed straight down through the bridge and against the delicate eggshell-like sounding box. To distribute the load and help the top plate withstand the downward component of string tension the viol makers glued to it a strip of wood running lengthwise down the middle. Whether by accident or by a stroke of genius, one of the earlier violinmakers moved the bar to one side so that one foot of the bridge rested above it. The strip, made of spruce, is now called the bass bar, since it is under the foot of the bridge on the side of the string of lowest tuning.

To support the other foot of the bridge there is placed approximately underneath it a vertical post, also made of spruce, called the sound post. It is carefully fitted and held in place between the front and back plates by friction. The acoustical function of the sound post has been a matter of debate for many years. The tone of a violin can be so greatly altered by small changes in the position, tightness and wood quality of the sound post that the French call it the soul (*l'âme*) of the instrument. Removing it altogether makes the violin sound rather like a guitar.

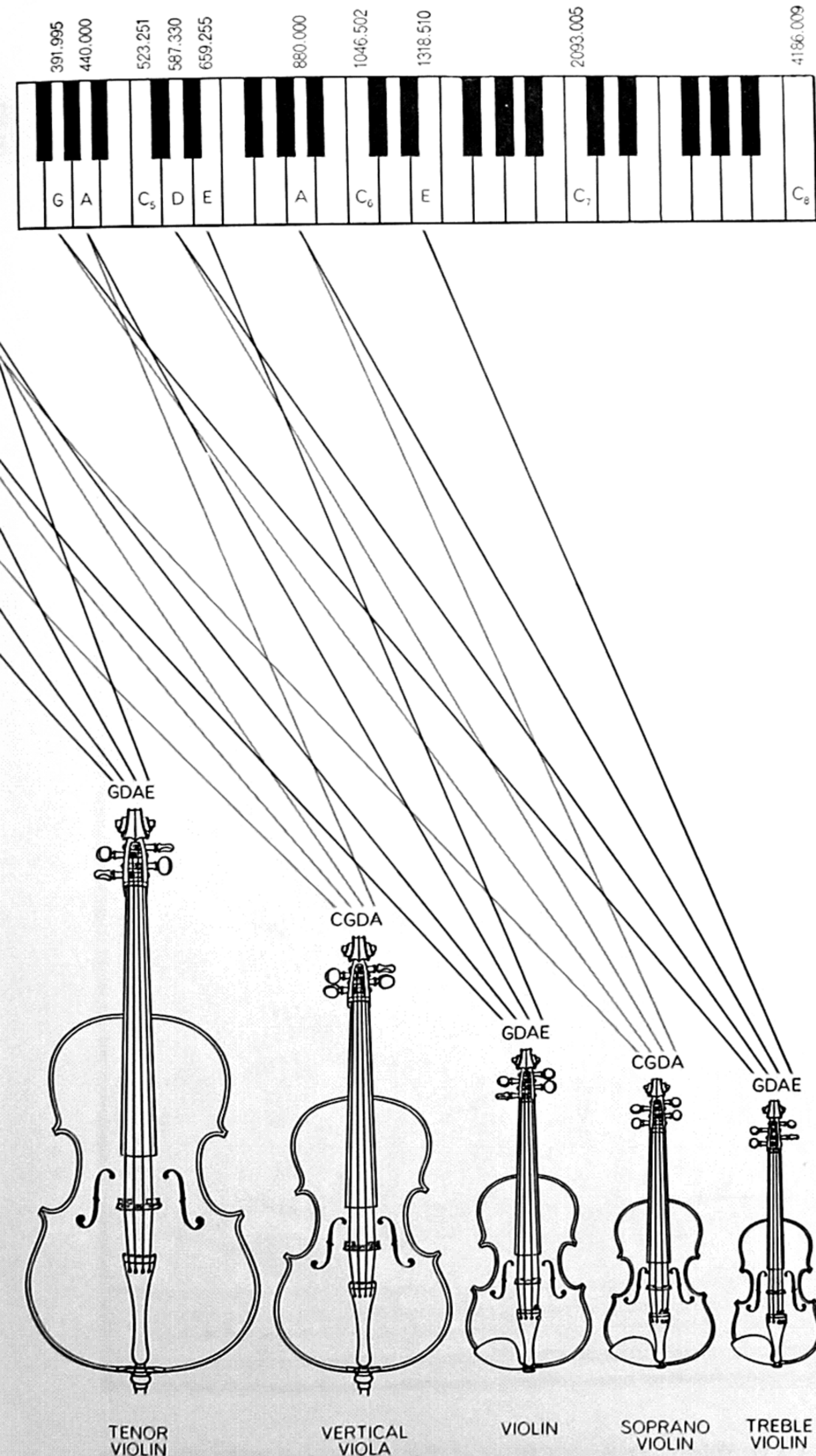
Although the modes of vibration of the plates exhibit great diversity throughout the frequency range, the bridge must always have some rocking motion to receive power from the string. In the important lower half of the range the sound post and the adjacent foot of the bridge have relatively little motion, thus providing in a sense a fulcrum that serves to transfer maximum travel to the bridge foot standing over the bass bar.

This too is getting a little ahead of the story, which begins when a bow is drawn across one or more of the strings of a violin. Vibrating strings have been stud-



NEW VIOLIN FAMILY will consist of eight instruments, twice the number of the present violin family. Of these the large bass and the treble violin are yet to be constructed; the small bass and the

tenor and soprano violins are newly designed instruments; the new cello and vertical viola are rescaled instruments; the violin is the only member of the original family to be included in the new



group unchanged (see illustration on page 773). The notes to which the four strings of each instrument are tuned can be read from the piano keyboard and the associated colored lines indicating the strings. The numbers at top show the frequencies in cycles per second.

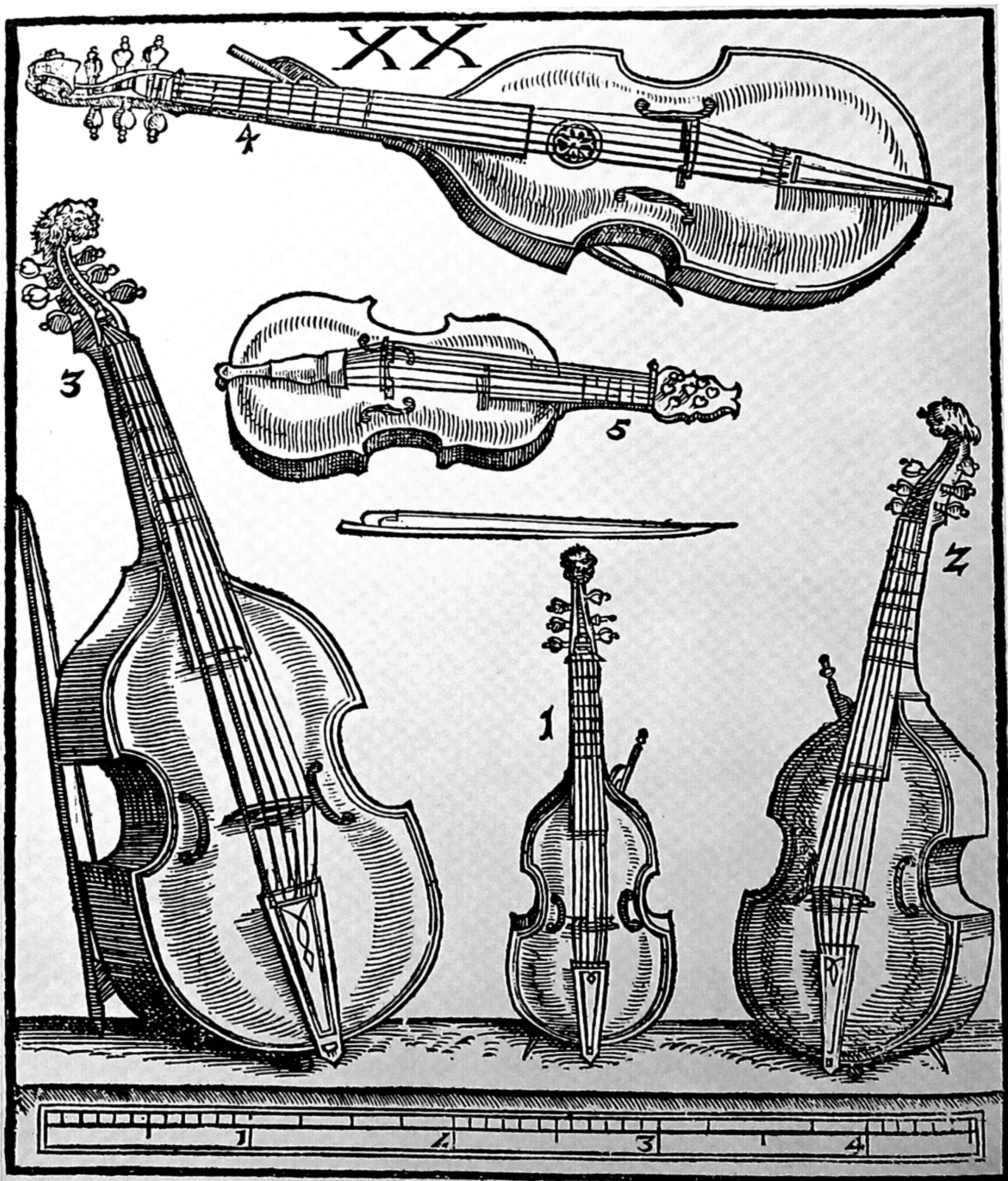
ied since the time of Pythagoras. An early 19th-century French physicist, Félix Savart, showed that the bowed string has a multitude of harmonics; then the great German physicist Hermann von Helmholtz elucidated the types of vibration that distinguish the bowed string from the plucked string. In our own century the Indian physicist Sir C. V. Raman made an exhaustive investigation of the vibration of bowed violin and cello strings. It would be fair to say that the reaction of a string to a bow is quite thoroughly understood.

In spite of the vigorous vibration of the moving string, the sound from the string alone would be all but inaudible. It has too little surface area to set an appreciable amount of air in motion. Trying to make music with an unamplified string would be like trying to fan oneself with a toothpick. What happens is that some portion of the energy supplied by the player to the bow—perhaps 5 to 10 per cent—is communicated to the wooden body of the violin through the complex motions of the bridge. (Of all the energy that the player feeds into the violin, 1 or 2 per cent emerges as sound. The rest goes off as heat.) The vibrations of the bowed string at any instant include dozens of energetic harmonics with amplitudes falling off as frequency increases. Each of the frequencies present shakes the wooden box—"forces" it to vibrate—at its particular rate. Obviously the amplitude of vibration depends on the strength, or amplitude, of the forcing vibration.

The Resonant Box

If this were all there was to it, matters would be simple, and all tones would be amplified equally. But the wooden structure itself has scores of frequencies at which it tends to vibrate naturally. The coincidence of such a frequency of resonance in the wood with the frequency of a string harmonic will result in an enhanced transfer of energy from string to box and a correspondingly greater amplification of that particular tone. Therefore the actual response of a violin to the playing of various notes is an enormously complex affair, but a good violinist must unconsciously and automatically deal with it and compensate for it every time he plays.

The scientific violinmaker is interested in all of these wood resonances, but he usually finds the resonance of lowest frequency an adequate guide during construction. This is called the main wood resonance. He is also interested in the lowest natural frequency (here only one



VIOL FAMILY, a somewhat earlier development than the violin family that eventually supplanted it, encompassed a large number of instruments of varying shapes and sizes. This drawing from the

Syntagma Musicum of Michael Praetorius (1619) shows three examples of the viola da gamba (1, 2, 3), a viola bastarda (4) and a viola da braccio (5). They have not all been drawn to the same scale.

seems to have any measurable importance) of the enclosed air space, called the main air resonance. Tests show that a good violin usually has its main wood resonance within a whole note of 440 cycles per second: the note A, to which the second highest string of the instrument is tuned.

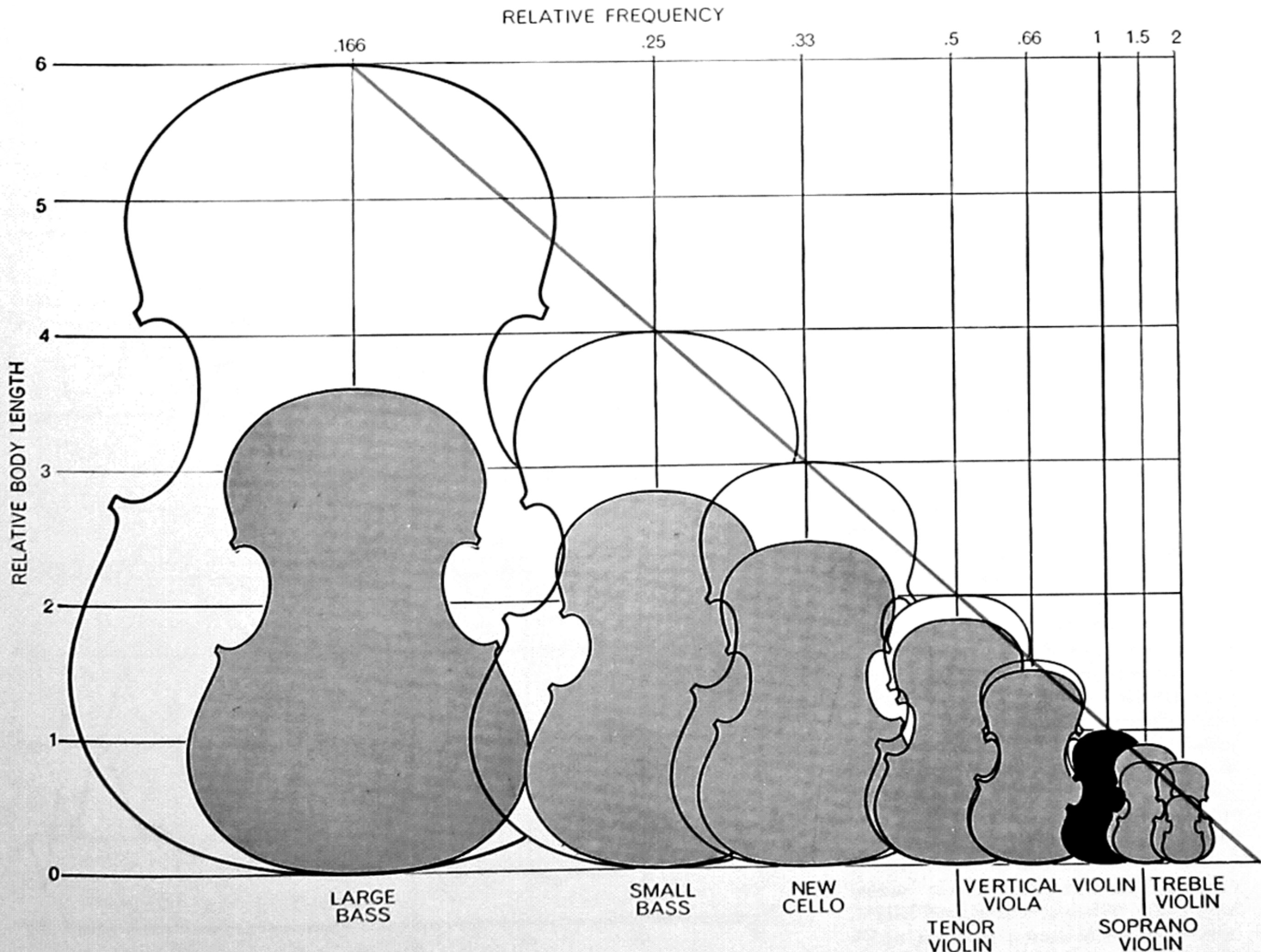
Some instruments have a "wolf note," almost always at the frequency of the main wood resonance. When this note is played on any string, the tone warbles unsteadily, often breaking by a whole octave somewhat as the voice of an adolescent boy does. The wolf note occurs when the string and the wood form a pair of mechanically coupled circuits; a beating action occurs because energy is cyclically shuttled back and forth between them. Violas and cellos are notori-

ously subject to wolf-note trouble. Even some of the finest cellos have bad wolf notes. It is possible to ameliorate this difficulty in a variety of ways, for example by tuning a length of string between the bridge and the tailpiece to the actual frequency of the wolf note. This absorbs enough energy to control the wolf. So far, however, the ideal method of control has not been found.

Inside the box of the violin is the air chamber, or resonating cavity, which communicates directly to the outside by means of the f-holes in the top of the instrument. As I have said, the enclosed air has so far been found to add measurable resonance to a range of tones surrounding one note on each instrument. The pitch of this main air resonance, or air tone, can be approximately located

by blowing across the f-hole, as one might blow across the top of an empty bottle. (When one f-hole is covered lightly, this pitch is lowered.)

The frequency of the air tone is controlled by the volume of air enclosed by the box of the instrument and the combined area of its f-hole openings. The larger the air volume, the lower the frequency; the larger the f-hole area, the higher the frequency. These two variables can be calculated roughly. I have found that to raise the resonance of the enclosed air a whole tone requires approximately a 20 per cent reduction in air volume or a 59 per cent increase in f-hole area. Anyone looking at the handsomely shaped f-holes of a violin can appreciate that it is not practical to try to raise the frequency of the air



THEORETICAL AND ACTUAL SIZES of new instruments are compared. According to theory an instrument of a given relative frequency would have to be of the size represented by the colored outlines. In practice the size is that represented by the light gray areas. The relative frequency of an instrument (e.g., treble violin)

is obtained by dividing the frequency of one of its strings (e.g., A) by the frequency of the corresponding string on the violin: 880 divided by 440 equals 2. The conventional viola, cello and double bass (not shown in illustration) have relative body lengths (using the violin as unity) of 1.17, 2.13 and 3.09 respectively.

resonance even a semitone by changing the size of the holes.

A number of workers—particularly Saunders, the late Hermann Backhaus of the Technische Hochschule in Karlsruhe, Hermann Meinel of Berlin, Gioacchino Pasqualini of Rome and E. Rohloff of the University of Greifswald—have developed methods of studying the resonances of violins. One of the most useful is the “loudness curve” originated by Saunders. It is also called the curve of total intensity, because it shows at each measured frequency the combined strengths of all the harmonics.

Loudness Curves

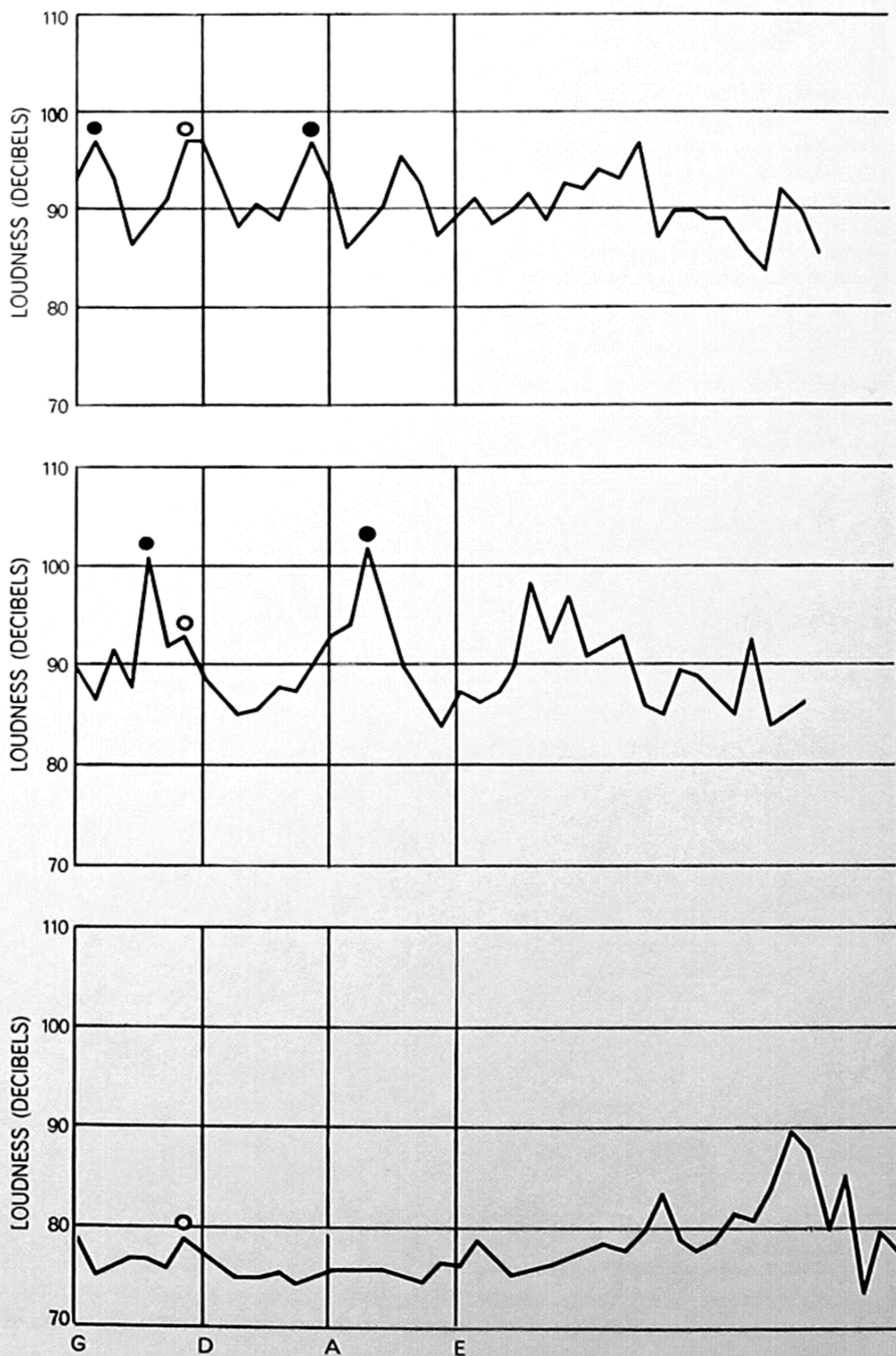
In making a loudness curve the violin is bowed normally, but without vibrato, at semitone intervals over its entire range to produce the loudest tone possible at each note. A General Radio sound-level meter, such as is used to measure levels of applause on television shows, records the loudness of each tone. It often comes as a shock to a musician to discover that his instrument is much louder at certain notes than at others. Try as he will he cannot possibly make them all register at an equally high level on the sound-level meter.

At the right are displayed loudness curves of a few violins, good and poor. In the good one the main wood resonance and the main air resonance fall approximately seven semitones, or a musical fifth, apart. (A fifth is the interval from “do” to “sol” on the diatonic scale. The frequency of each note is in the ratio of three to two for the note below it.) In some poor instruments the main wood and air resonances may be as much as 12 semitones, or an octave, apart (frequency ratio, two to one), giving two areas of strong resonance with a wide range of weak response between. The curve for one poor instrument shows only one area of strong resonance—the air resonance—with the wood contributing virtually nothing in the way of resonant reinforcement. A \$5 violin with a curve almost as bad was used by Saunders for some time as his “standard” of badness. When I took the wretched thing apart and balanced it for good tone production, it showed an overall increase of loudness and an even spacing of peaks. At this point it was named Pygmalion. When it was played behind a screen in alternation with an excellent Cremona violin, the two were voted equal in tone by a college music department audience. In fairness it should be added that the skilled musician playing behind the screen was never in any doubt

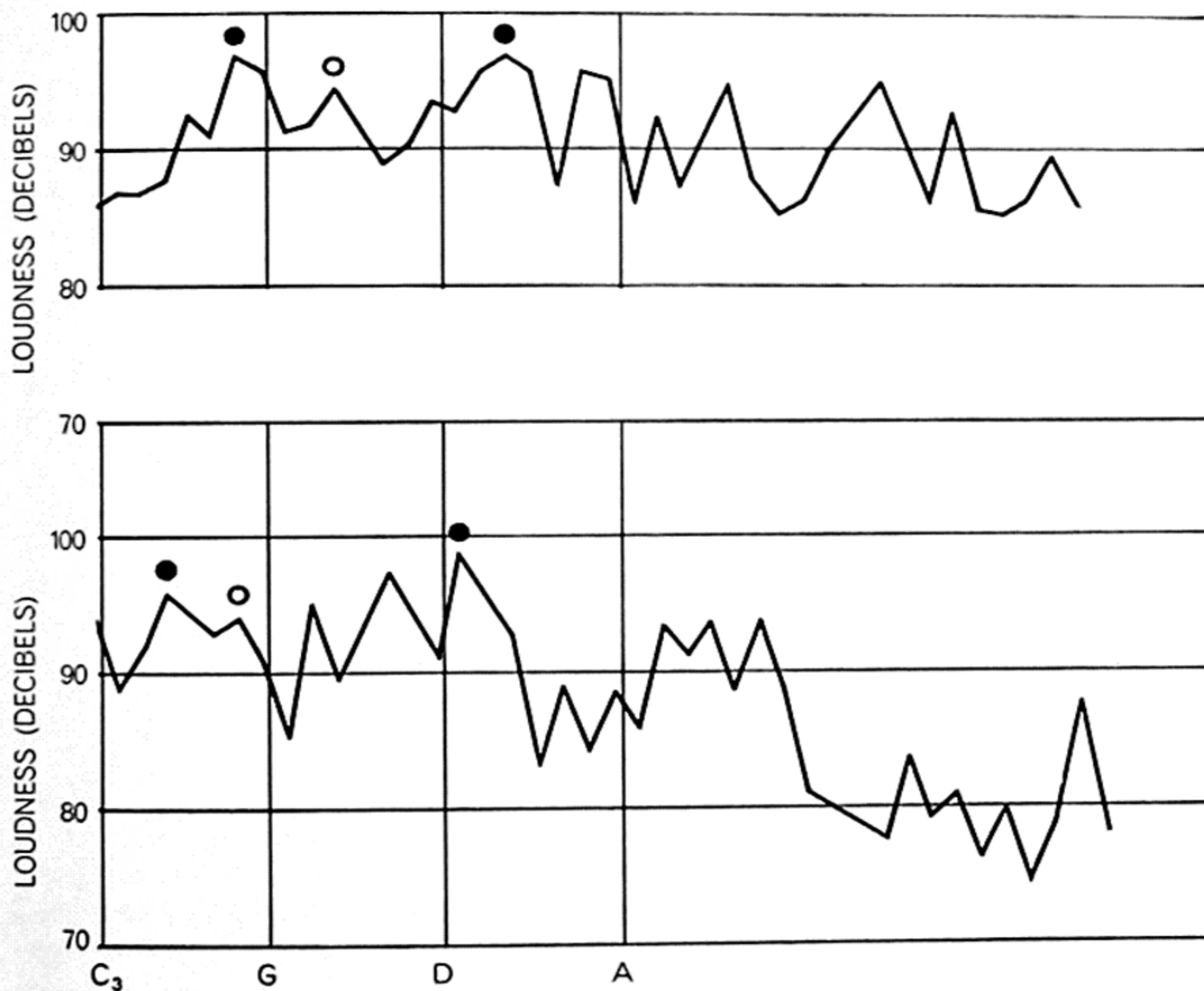
as to which was the superior instrument.

An octave below the main wood resonance there is almost always another strong peak of loudness that we label “wood prime.” It can be called a subharmonic. It is well known in acoustics that if one harmonic of a complex tone

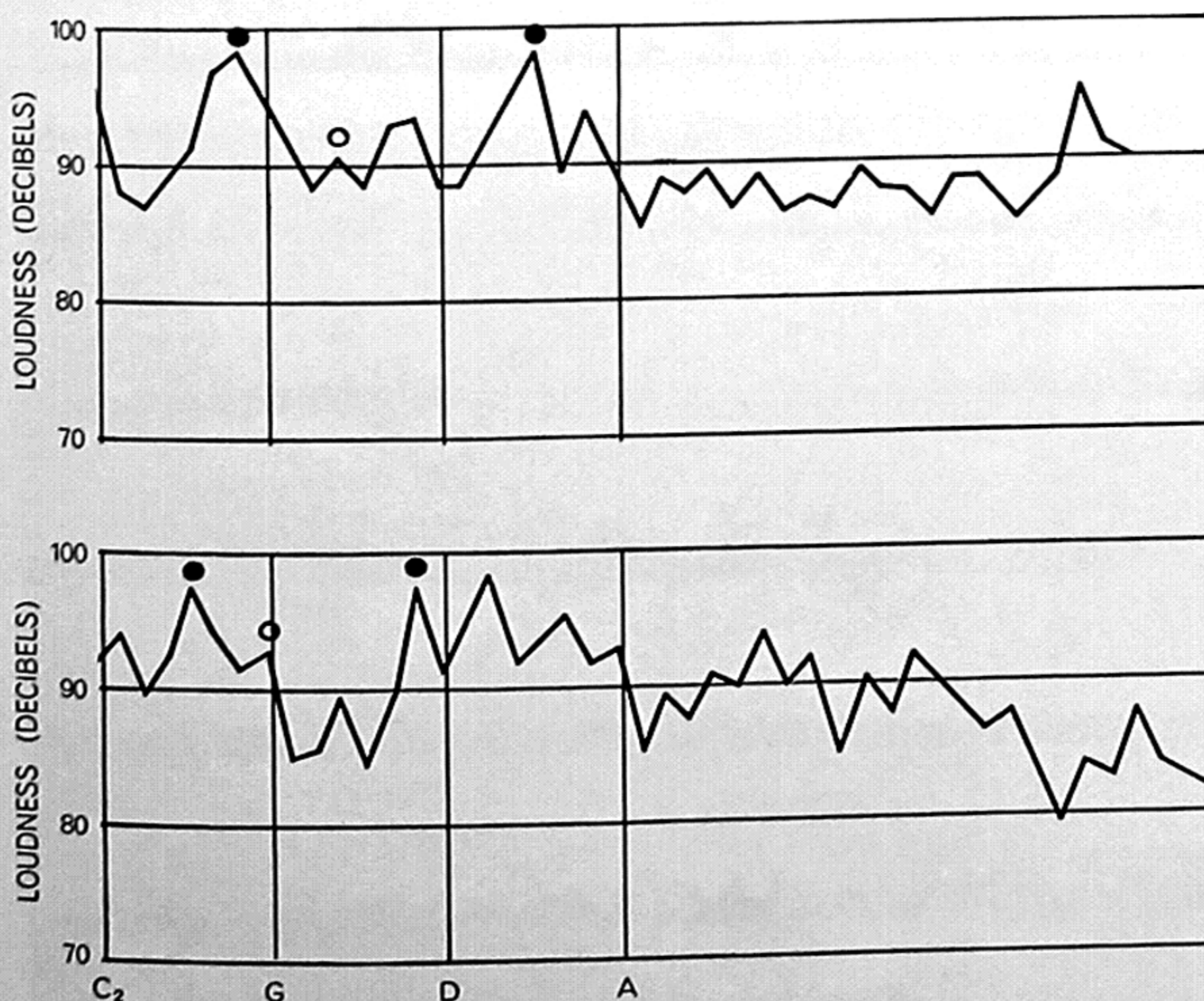
is strengthened, the ear will hear an increase in loudness of the note as a whole with a slight change in quality but no change in pitch. By this process the wood peak is strengthened by the tone of the main wood resonance an octave above. The subharmonic of the main wood reso-



VIOLIN “LOUDNESS CURVES” compare maximum sound levels produced at semitone intervals by a good 1713 Stradivarius (*top*), a poor 250-year-old violin of doubtful origin (*middle*) and a poorer, somewhat older instrument credited to P. Guarneri (*bottom*). Only the first shows desirable spacing and strength of wood (*black dot*), “wood prime” (*gray dot*) and air resonances (*open circle*). Letters at bottom indicate tuning of open strings.



VIOLA LOUDNESS CURVES compare the responses of a conventional (*top*) and a vertical viola (*bottom*), both made by the author. The convention of dots and colored lines used in this illustration and that below is the same as in the illustration on preceding page.



CELLO LOUDNESS CURVES compare the responses of a conventional (*top*) and a new cello (*bottom*), both made by the author. Note that the main wood and air resonances of the cello are near the two open middle strings. The loudness curves on this and preceding page are based on tests performed by Frederick A. Saunders of Harvard University.

nance benefits the lower tones of the violin, viola and cello.

The subharmonic of the main air resonance does not show on curves of conventional instruments because it falls below the bottom notes of the instruments. Spacing the main wood and air resonances about a half-octave apart spreads these peaks so that the air-tone peak falls nicely in the middle of the octave between the wood resonance and its subharmonic. In hundreds of tests of violins, violas and cellos this arrangement of wood and air resonances emerges as one of the characteristics of the good instruments.

Experimental Instruments

I have built a series of experimental violins and violas to test the effect of moving the frequencies of the main resonances up or down the scale. In a pair of violas of similar pattern with identical f-holes, one was made with sides half an inch high to decrease the air volume; the other had sides two inches high, giving a large air volume. Normally the sides of a viola are about 1½ inches high, and the air tone is found to be in the range from B to B flat (233 cycles per second) on the G string. In the viola with the smaller air volume the air tone, as expected, moved up the scale to D sharp (300 cycles per second). In the one with the larger air volume the air tone moved downscale near A (220 cycles per second).

In both of these altered violas the normally strong tones of the B to B flat on the G string were missing, because the air resonance was no longer there to reinforce them. Musicians playing the instruments discovered interesting features. Neither one was suitable for playing the two-violin quintets composed by Mozart. The composer had written so well for the normally strong tones of the viola that the outstanding parts lacked their full expressive qualities when the experimental instruments were used. The strong resonance of the air tone was not where musicians expected to find strength, nor where Mozart had counted on it.

The most interesting feature of the two violas was that the thin, shallow instrument had a full, rich tone and a particularly strong, low C string, where the normal viola is notably weak. This was because the air tone had been shifted upscale enough so that its subharmonic came into useful range near 150 cycles per second on the low C string.

The thick viola with the two-inch ribs,

on the other hand, had a thin tone, and the lower range of its C string was weak, partly because the air tone had been moved from its normal position. Many musicians playing the two violas in alternation have remarked with astonishment at the full, rich tone of the thin one with the small air volume. Ribs half an inch high are structurally not very practical, but application of the principles involved has made possible the construction of good small violas.

In studying the resonances of violins I have discovered that in the best violins the main wood and air resonances invariably fall within a semitone or two of the frequency of the two open middle strings, the wood resonance corresponding to the higher-tuned string. When the early violinmakers hit on this arrangement, the muses must have been smiling. It is, quite simply, the way in which most good violins have been made ever since.

This is not true of the viola and cello. In these instruments as they are now

built the wood and air resonances fall three to four semitones higher with respect to the frequencies of the open middle strings than they do in the violin. The reason is simple enough: the viola and cello are built smaller than optimum size to make them a convenient playing size. As a result the resonances are too far above the lower notes of the instruments, and these suffer in strength and quality. I shall have more to say about this matter later.

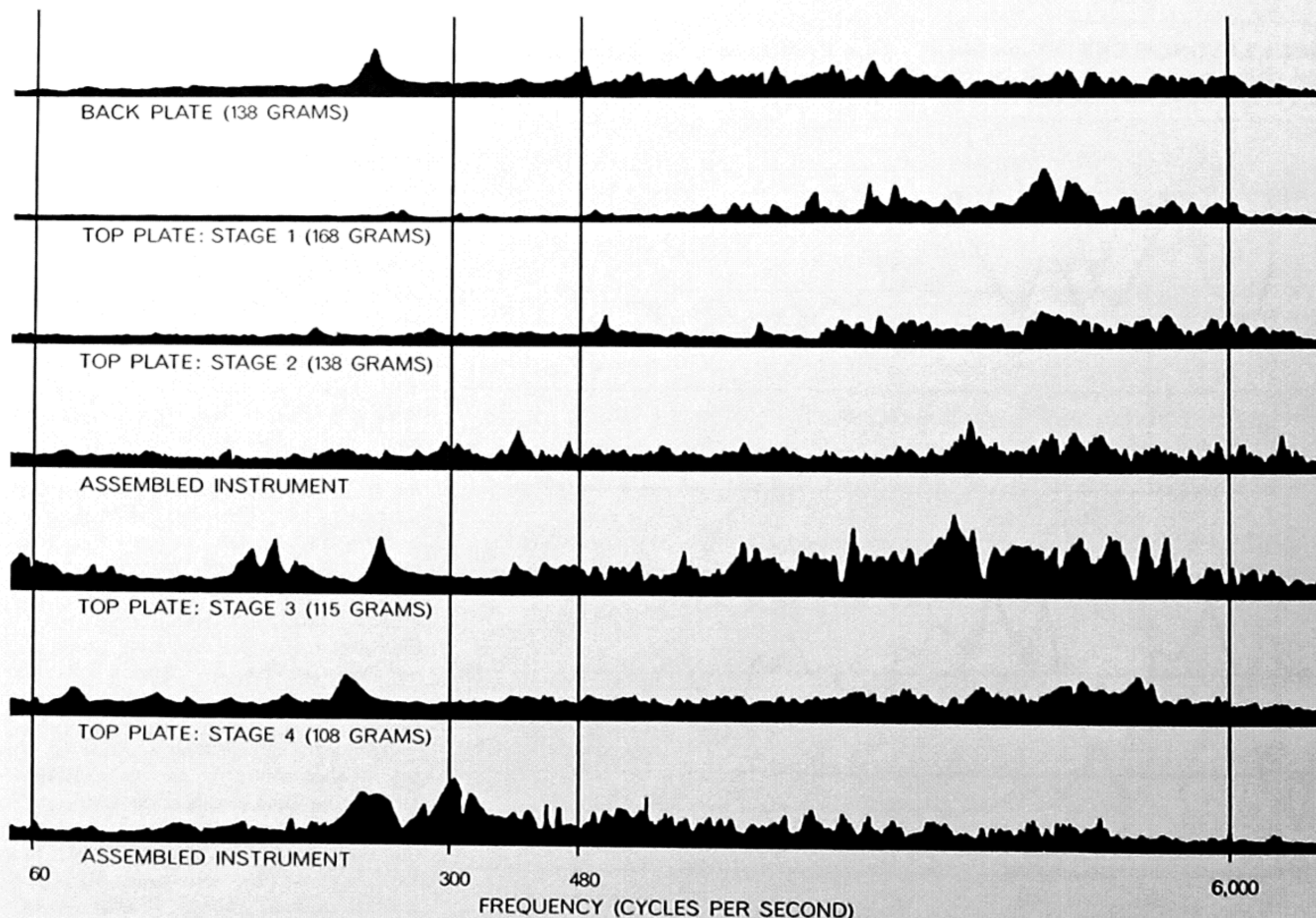
Tap Tones

At the moment I should like to consider a different problem. Assuming one knows what the violin should be like when it is finished—where its resonances should fall and so on—how are these aims achieved in the process of construction? How does one make it come out the way one wants it? In addition to careful workmanship and accurate measurements the traditional method of the violinmaker has been to listen to the

“tap tones” of the front and back plates.

In the final thinning and graduating of the top and back plates of a violin, the maker traditionally holds the plate near one end in his thumb and forefinger, taps it at various points with a knuckle and listens carefully to determine the pitch of the sounds he hears. These sounds are called the tap tones of the plates. The ability to judge the proper relation of tap tones of the free top and back plates in this manner is an important part of the art of violinmaking. With the ear alone it is extremely difficult to make out the frequencies of these tones, particularly in the case of the top plate, where the complicated structure with f-holes and bass bar creates at least two, and sometimes as many as five, strong natural resonances below 600 cycles per second.

Saunders and I, together with Alvin Hopping of Lake Hopatcong, N.J., have developed a method that makes it possible to determine the tap-tone frequencies in a free plate with considerable



FREQUENCY-RESPONSE CURVES of top and back plates of a viola and of the assembled instrument at various stages are depicted. Although the tests run from 20 to 20,000 cycles per second, most of the response to the magnetic driver used to vibrate the

wood falls in the range of 60 to roughly 10,000 cycles per second. The four frequencies indicated here are those at which checks were made to ensure that the recording film was synchronized with the audio-generator. Height of peaks represents amplitude of response.



CHLADNI PATTERNS discussed in the text were made on a brass plate by Saunders at frequencies of 260, 340, 435, 520, 780 and 1,600

cycles per second. The plate is supported horizontally by a bolt at the center of its upper half; bottom end rests on a padded block.

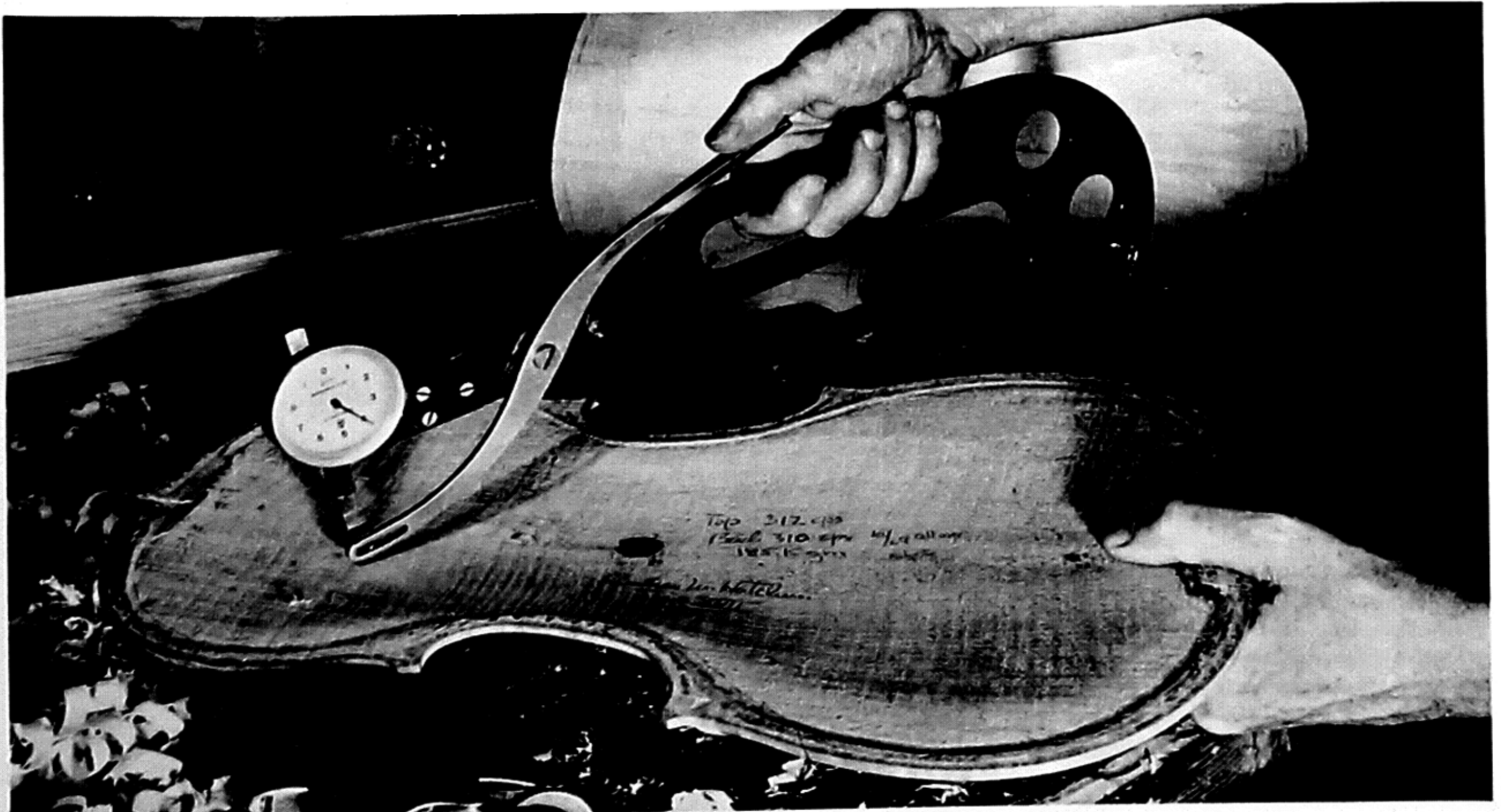
accuracy. First we cut a flat brass plate in the shape of the violin plate. We dust it with powder and bow it at various points around the edge to set up different modes of vibration. Where the plate vibrates, the powder is bounced away, piling up along the nodal lines where there is no vibration. From these "Chladni" patterns on the brass [see illustration above] we are able to predict where a principal nodal point in the frequency test range will fall on the mid-line of a real violin plate. Since clamping on a nodal point does not affect the vibration pattern, we then clamp the violin plate at this point and set it into vibration at its exact center

by means of a magnetic driver, activated by an audio-frequency generator, that can be varied from 20 to 20,000 cycles per second. The response of the wood plate to the input signal, which has variable frequency but constant amplitude, is picked up by a microphone and fed to an oscilloscope or a sound-level meter. The amplitude and frequency of the points of greatest response can be recorded manually. Better still, a "photostrip" can be made by pulling a film across the oscilloscope face at a speed synchronized with the sweep of the audio-frequency generator.

Once we had established the testing procedure we could address ourselves to

a question that had been worried for several hundred years and that had been answered in a number of different ways: What sounds should the top and back plates of an instrument produce before they are joined?

In 1840 Savart reported that "a top of spruce and a back of maple tuned alike produced an instrument with a bad, weak tone." He took the plates off a number of Stradivarius and Guarnerius violins (imagine!) and tested them, finding that the tap tones varied "between C sharp 3 and D3 (in the octave above middle C) for the top, and D3 and D sharp 3 for the back, always one tone or one semitone difference, the



MEASURING PLATE THICKNESS makes it possible to determine where thinning can be done while maintaining a fairly uniform

pattern of thickness. The measuring device, or caliper, consists of a dial gauge attached to one arm of an extended metal U.

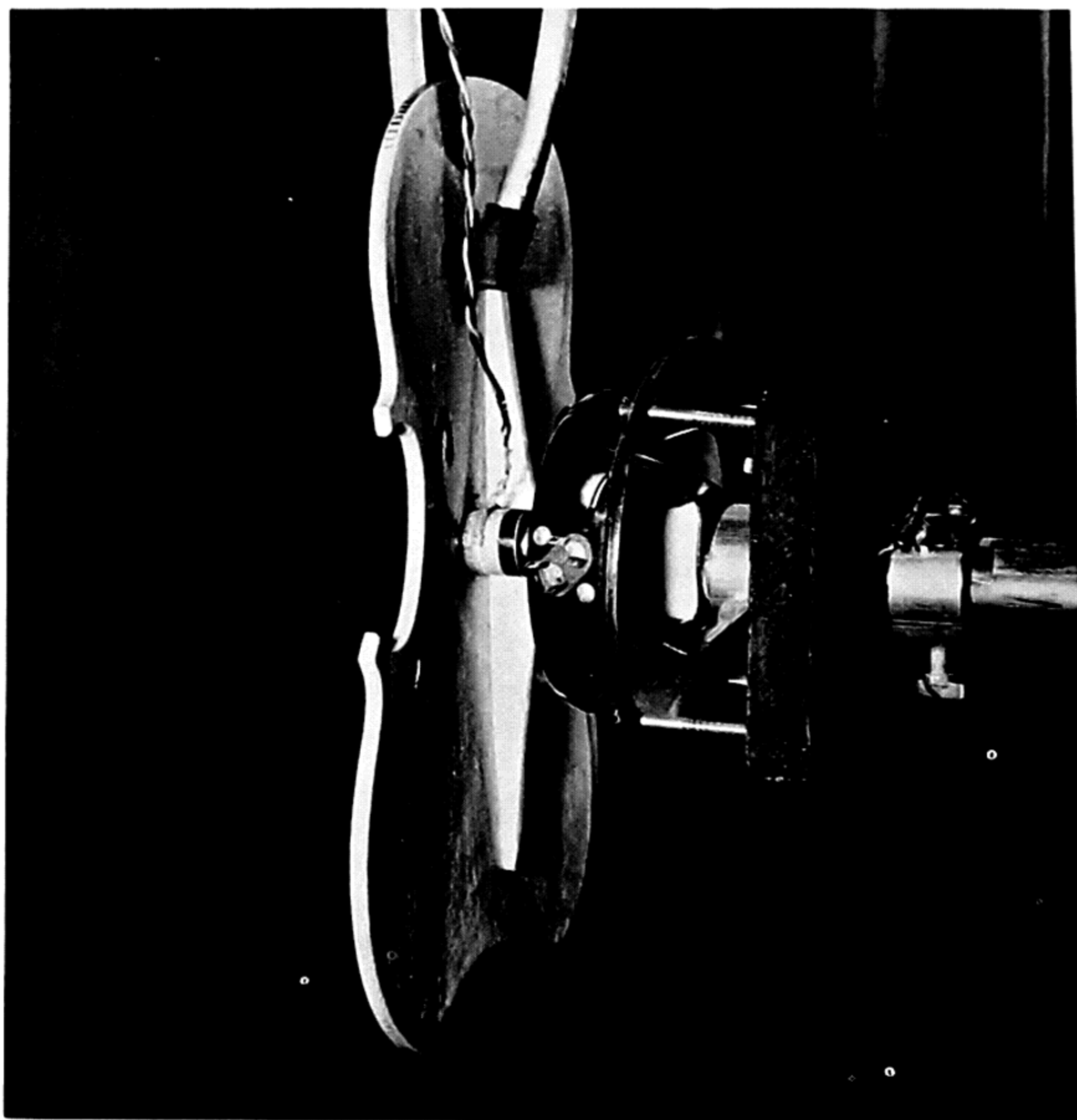
back being higher than the top." Some violinmakers have held that the back should be a tone lower than the top; others, that the plates should be tuned to the same frequency.

My own findings are as follows: In the range of 120 to 600 cycles per second there may be one, two and possibly three peaks in the back plate and perhaps two or three more than that in the front. When the peaks of the front plate alternate with those of the back and the adjacent peaks are within about a semitone of one another, I get a good instrument. When the peaks coincide or are more than a tone apart, I get a bad one. Moreover, an average of the frequencies of the tap tones from front and back turns out to be just about seven semitones below the main wood frequency of the finished violin.

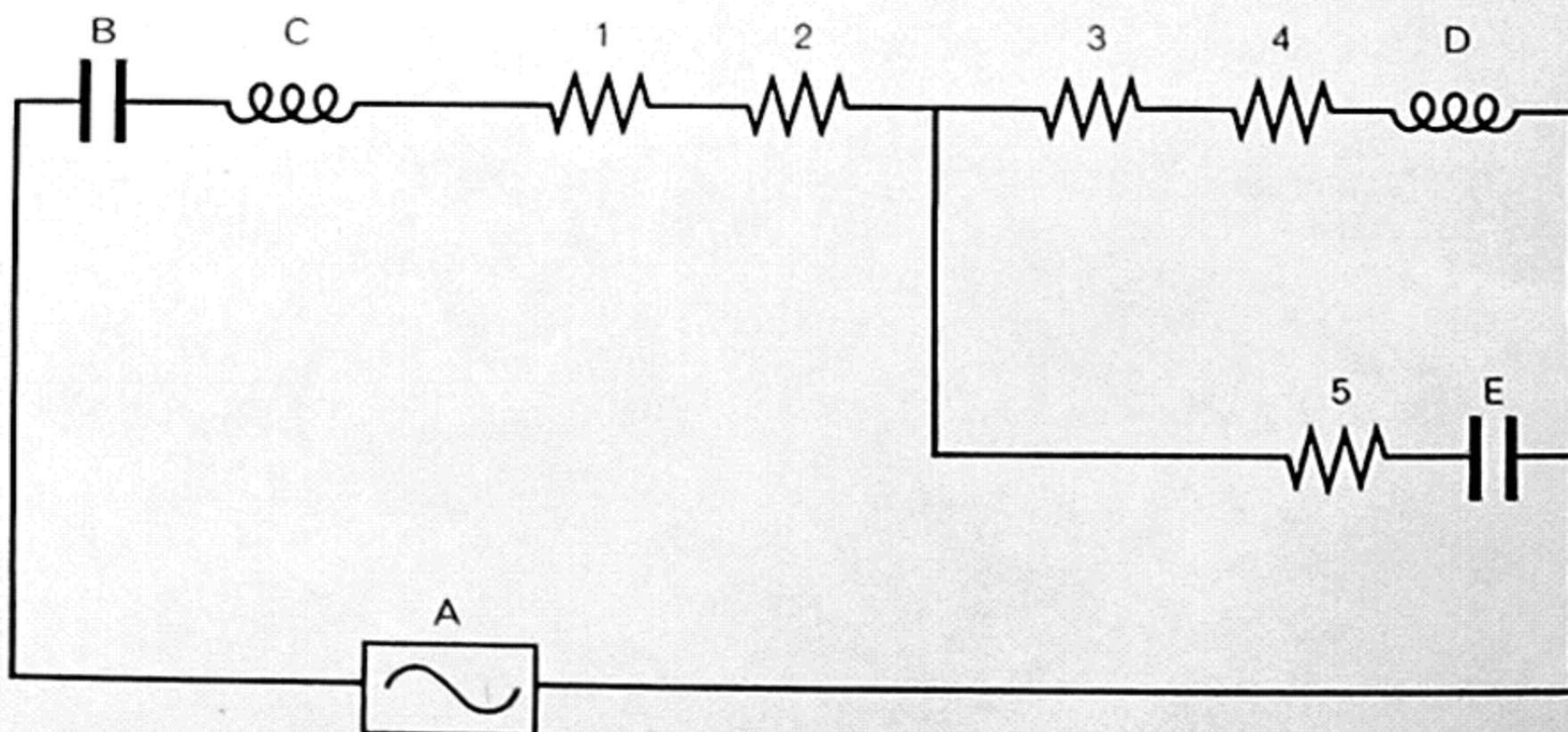
These conclusions are drawn from more than 400 photostrips of top and back plates of 35 instruments in the process of construction. After the plates were tested the instruments were assembled and then judged for tone quality by three criteria: (1) loudness test, (2) photostrips of the completed instrument and (3) actual playing by professionals. Then one plate, usually the top one, was removed and thinned, tested again and the instrument assembled for reappraisal. The back plate was thinned only when the top plate became so thin that it could no longer support string tension with safety. The entire thinning and testing process was sometimes repeated four or more times until each violin, viola or cello was judged to be good. So far I have spent six years on the program.

With our tap-tone test it is possible to follow the position of the main wood vibrations as they drift to lower frequencies when the wood is thinned and becomes more flexible. With a little practice one learns how to remove a few grams of wood from certain areas with a scraper or small plane and to estimate that the plate peaks (strong natural resonances) will move downscale, say 10 cycles per second. In some cases such a shift can make the difference between a good and a poor instrument.

As a kind of acid test of the theory I made a cello with the plate peaks matching; this is of course exactly wrong. During the next two years I gave the cello to several different cellists to play. All of them admired the workmanship and tried to be complimentary about the tone and playing qualities. The more forthright of them said that the tone was harsh and gritty in spots and weak in others and that the instru-



MAGNETIC DRIVER used in frequency-response tests is placed at the exact center of a top plate. The wires leading from the driver are connected to an audio-frequency generator that activates the driver over a range of frequencies from 20 to 20,000 cycles per second.



SIMPLIFIED ELECTRICAL CIRCUIT shows the nature of the two main resonances discussed in text. The current from a constant-amplitude alternating-current generator (A) is analogous to the force applied by a given string to the bridge; this force is proportional to the string tension and amplitude of string vibration. The first capacitor (B) is analogous to a stiffness associated with the elasticity and dimensions of the wood; the first inductance coil (C) is analogous to a mass moving with the velocity of the bridge-string contact and having a kinetic energy equal to that in the wood. In instruments of the violin family the stiffness and mass of the wood largely determine its over-all response and the frequency of the main wood resonance. The main air resonance is determined largely by opposition of the air to compression when the f-holes are closed (E) and the mass of the air near the open f-holes (D). The five resistors in the circuit represent mechanical or acoustical resistances.

ment was particularly hard to play softly.

Finally I took the plates off, tested them again and removed about 10 grams of wood from the edges of the top plate so that the peaks of the top alternated with those of the back. In this condition Mischa Schneider played the cello in a concert by the Budapest String Quartet and pronounced it to be *magnifico*.

The greatest difficulty with the tap-tone test on a finished instrument is that both the top and back plates must be off at the same time so that they can be tested under the same conditions and without the complication of drift in the measuring equipment. The removal of both plates is a touchy operation even for an expert. With the help of several co-operative violinmakers, however, we have been able to test the plates of a few good old violins. More such tests are needed for definitive comparisons.

New Violins

It has been hundreds of years since the violin won its battle with the viols. The victory was not an unmitigated blessing. The variability of the shape of the viols, and particularly their flat back plate without a complicated set of resonances, meant that the instruments could be built in a variety of sizes that easily covered the entire range of pitch represented by the piano keyboard. On the other hand, the violin family leaves substantial gaps in coverage and, as has already been pointed out, its two deeper-voiced members do not have optimum musical characteristics.

Buried in private collections and museums there is a neglected but rich repertoire of polyphonic string music from the Renaissance period written for viols. Their characteristically thin and nasal, but uniform and distinctive, timbre blended well with the clavichord and cembalo, which were played at the courts of Renaissance nobles. Many of these gentlemen kept a chest of viols, usually consisting of six instruments, two each of the treble, alto and tenor sizes.

For contemporary performance of the viol repertoire, however, the old instruments are unsuitable. They do not have the variety of timbre that the violin has taught the modern ear to expect, and they do not have nearly the power to satisfy the requirements of a concert hall of even moderate size. On the other hand, the present family of violin, viola and cello have too much inequality in timbre and too great gaps in pitch to play the music as it was written.

The need for new instruments of the

violin type has been considered by musicians and violinmakers for many years. The present work of developing the new instruments was initiated when Henry Brant, the composer-in-residence at Bennington College in Vermont, came to us with the problem. Brant felt that modern musicians, faced with the need to find an expressive language appropriate to the present day, are ever seeking to extend the powers of the bowed string instruments. The violin family remains the composer's most eloquent and expressive vehicle among all the instruments so far devised in Western music, but its members have been essentially unchanged for 200 years. More and more the need is being recognized for a gamut of graduated instruments of the violin type, with each member well enough developed to meet the test of solo as well as ensemble playing.

Changing the classical dimensions of the violin to create instruments of varied sizes and tunings has been tried many times without success. Now the necessary knowledge is at hand. I have indicated that the variables in the design of the violin are close to optimum. The object is to keep the two main resonances on the two open middle strings in spite of changes in size and tuning. It can readily be appreciated that it is no mean task to arrive at the correct proportions among physical variables—size, thickness and stiffness of wood, tightness of stringing and so on—that will produce the desired result in the resonance. For me it has meant years of literal cut and try, but with the help of scaling theory I am now close to having a set of empirical rules for making a genuinely complete family of instruments of the violin type. In doing this I have drawn heavily on the knowledge gained by other violinmakers who have tackled the same problem but without success because they did not have the benefit of modern acoustical physics. I have already built revised versions of the viola and cello, enlarging them somewhat to bring the resonances down to the frequencies of the open middle strings. As a result my viola has to have a peg at the bottom, like a cello, and is played between the knees. In addition I have added two new instruments to the family (one replaces the bass). This past January the six scaled members of the violin family were tried out at an informal concert, which a number of professional musicians found interesting and challenging as well as aesthetically pleasing. The smallest and the largest of the new instruments have not yet been finished and are giving the most trouble. Although scaling theory

tells us what to do, we are up against the limits of available materials and the human physique. For the smallest instrument, which is tuned an octave above the violin, material of sufficient tensile strength for strings is the major problem. Few materials have the strength to vibrate within the requisite range of frequencies and still provide strings long enough to allow the player to finger consecutive semitones. In the largest instrument the designer faces the mechanical problem of making it possible for the musician to bow and finger simultaneously.

Other violinmakers have experimented with instrument size. In the 19th century Jean Baptiste Vuillaume introduced a new model of the viola with an exceptionally large air volume, constructed on the scientific principles of Savart. He also developed a huge double bass, known as the *octobasse*, that was tuned by means of levers. Fred Dautrich of Torrington, Conn., spent much of his time during the 1920's and 1930's working on a graded series of instruments of the violin type that he called the *vilonia*, the *vilon* and the *vilono*. I have been fortunate enough to obtain a set of these. They are of such excellent workmanship and proportions that it has been possible to modify them slightly by applying scaling theory and adapt them to our present series of instruments.

In the past few years J. C. Schelleng, formerly of the Bell Telephone Laboratories, has been studying the violin as a circuit, one of the standard techniques of acoustics in which the various mechanically vibrating parts are treated in a manner analogous to the elements of an electrical circuit. Although the violin is exceedingly complicated, it possesses many simplicities not usually recognized. These, along with the fundamental physics of the instrument, permit the definition of "circuit elements" and lead to relations difficult to find empirically. This circuit concept is already being of great help in perfecting the new instruments, defining such problems as string tension, the mass of the box and the stiffness of the plates.

To sum up, I believe that, without ignoring the precious heritage of centuries, the violinmaker should become more conscious of the science of his instrument, and that the acoustical physicist should see that here is a real challenge to his discipline. We really ought to learn how to make consistently better instruments than the old masters did. If that challenge cannot be fulfilled, we should at the very least find out the reasons for our limitations.

The Author

CARLEEN MALEY HUTCHINS has been designing and constructing violas and other stringed instruments of the violin family for the past 15 years. Her first step toward a career as a luthier came in 1942, when, as she describes it, "I bought an inexpensive weak-toned viola because my musical friends complained that the trumpet I had played was too loud in chamber music, as well as out of tune with the strings—and besides they needed a viola." The viola was unsatisfactory and Mrs. Hutchins turned for help to her uncle, William Harvey Fletcher (not the well-known acoustical physicist Harvey Fletcher), who had made many violins himself. Fletcher declined to try his hand at making a viola and instead directed his niece to a dealer who could supply her with the requisite books, blueprints and wood. In 1947 Mrs. Hutchins took a leave of absence from the Brearley School in New York, where she had taught science since 1937, to have her first child. Encouraged by her husband, she also embarked on the task of making her first viola, a job that took two years. Mrs. Hutchins has produced 55 instruments, selling some of them to help pay for further research. A graduate of Cornell University, where she studied entomology, Mrs. Hutchins retired from

teaching in 1949. For the past 12 years she has collaborated with Frederick A. Saunders of Harvard University in the study of the acoustics of the instruments of the violin family. At present she is continuing this work under her second Guggenheim Fellowship.

Bibliography

THE MECHANICAL ACTION OF INSTRUMENTS OF THE VIOLIN FAMILY. F. A. Saunders in *The Journal of the Acoustical Society of America*, Vol. 17, No. 3, pages 169–186; January, 1946.

THE MECHANICAL ACTION OF VIOLINS. F. A. Saunders in *The Journal of the Acoustical Society of America*, Vol. 9, No. 2, pages 81–98; October, 1937.

MISURA DELL'ATTRITO INTERNO E DELLE COSTANTI ELASTICHE DEL LEGNO. I. Barducci and G. Pasqualini in *Nuovo Cimento*, Vol. 5, No. 5, pages 416–446; October 1, 1948.

REGARDING THE SOUND QUALITY OF VIOLINS AND A SCIENTIFIC BASIS FOR VIOLIN CONSTRUCTION. H. Meinel in *The Journal of the Acoustical Society of America*, Vol. 29, No. 7, pages 817–822; July, 1957.

SUBHARMONICS AND PLATE TAP TONES IN VIOLIN ACOUSTICS. Carleen M. Hutchins, Alvin S. Hopping and Frederick A. Saunders in *The Journal of the Acoustical Society of America*, Vol. 32, No. 11, pages 1443–1449; November, 1960.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound **SCIENTIFIC AMERICAN** Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

RESONANCE PARTICLES

by R. D. Hill

Most of the 32 fundamental particles of matter decay rather quickly. There are still other particles that decay even more quickly. It now seems that the latter are "resonant" associations of other particles.

When is a particle "fundamental"? The fact that most subatomic particles decay quite quickly into other particles has made this a perennial question of physics. Now the difficulty is compounded by a new group of "particles" that are even more evanescent than the particles known earlier.

The list of "old" particles, which are generally considered fundamental and which I shall call Type I, has been static for some time [see illustration on page 784]. It still consists only of those particles mentioned as discovered or predicted in an article that appeared in SCIENTIFIC AMERICAN more than five years ago [see "Elementary Particles," by Murray Gell-Mann and E. P. Rosenbaum; SCIENTIFIC AMERICAN Offprint 213].

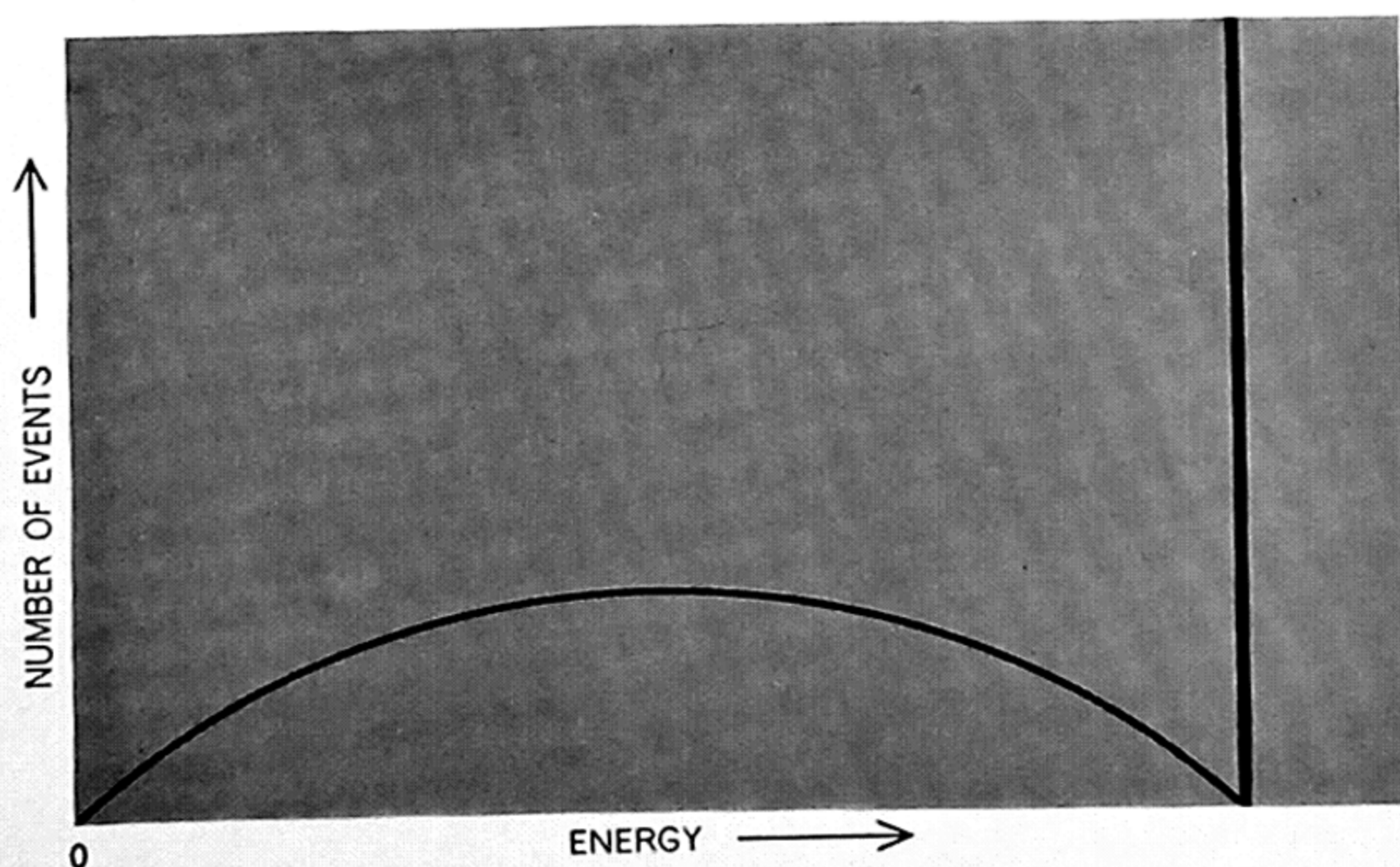
Apart from the inherently stable particles, most of the Type I particles decay in about a ten-billionth of a second. This lifetime gives them a chance to move measurable distances in a detector such as a bubble chamber. The new particles, which I call Type II, decay in about the time it takes for light to move a distance equal to a few diameters of an atomic nucleus. Their lifetimes are measured in hundred-thousandths of a billion-billionth of a second (10^{-23} second)—far too short a time to leave visible tracks or to be observed directly in any way. Their existence can only be inferred by studying the Type I particle products of their disintegrations. The question is: Were they ever autonomous particles or were they merely a group of separate pieces that moved together for a short time before flying apart? Physicists have avoided the question by calling the particles "resonances," implying that they may indeed have been temporary associations of other particles.

How is their existence detected at all? To help the reader understand the way in which they were found I shall begin by describing a fanciful experiment in classical physics, that is, the kind of physics that obtained before the introduction of quanta and relativity. Like all classical explanations of quantum and relativistic events, the analogy is far from perfect. It should nonetheless serve to provide a rough idea of what is involved.

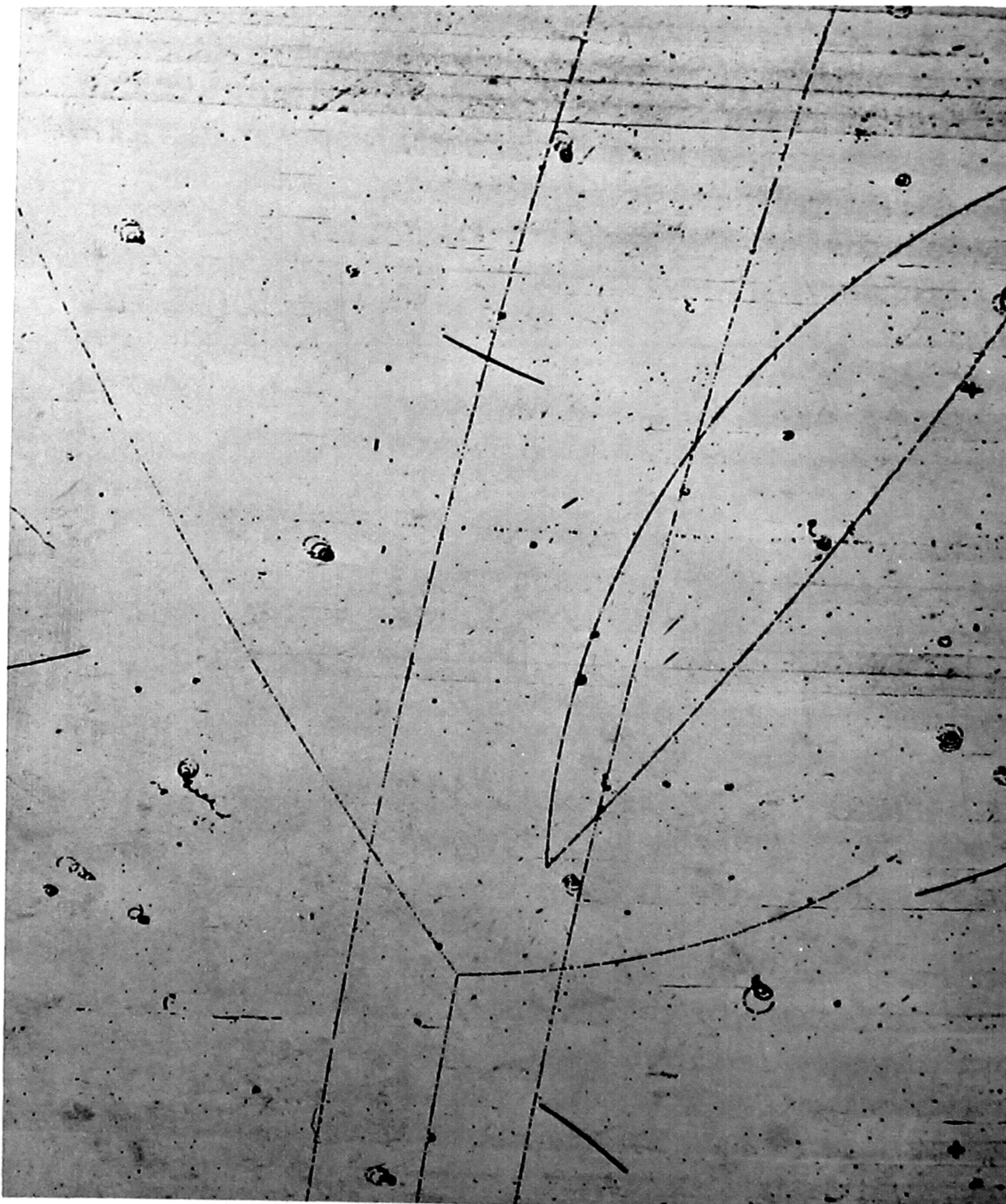
Suppose an odd kind of artillery shell is made by gluing together three pieces of strong glass around an explosive charge. One of the pieces is red; the other two are perfectly transparent and therefore invisible. The shell is fired, moving, let us suppose, on a straight

trajectory at a constant speed. As observers we move parallel to the shell in an airplane traveling at the same speed. When the shell explodes, it breaks into its three component pieces. The center of mass of these pieces continues to move in its original direction at its original speed, according to the law of the conservation of momentum. We too continue to move in this direction and at this speed. Therefore it is as if the shell and we were both stationary when the shell exploded. We are observing in the "center-of-mass system" of the exploding fragments.

Since two of the three pieces are invisible, after the explosion we can see only the red piece. We measure its speed with respect to our own position. If we

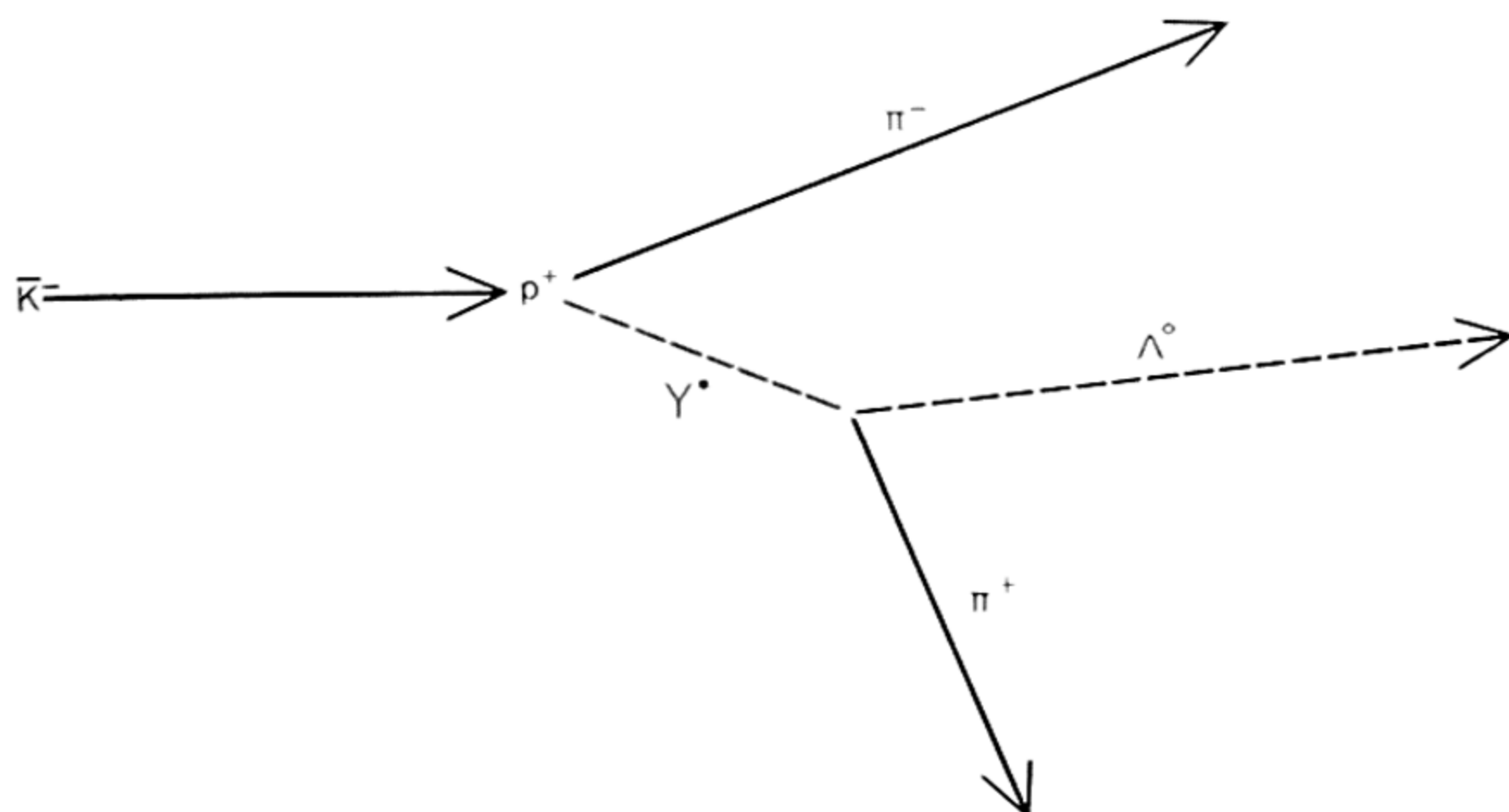


DISTRIBUTION OF ENERGY to one of three pieces of an exploding shell (in the imaginary experiment described in the text) over a number of events can be represented by a smooth curve. But if two of the three pieces always stick together, the third piece will, in theory, always receive the same amount of energy; its energy curve becomes a straight line.



Y^* RESONANCE PARTICLE and a negative pion (π^- from O in drawing at right) are produced in this bubble-chamber collision between a negative K meson (\bar{K}^-) and a proton at O . The resonance

particle disintegrates, before it can leave a track, into a neutral lambda particle, which leaves no track (*broken line*), and a positive pion (π^+). The lambda decays into a proton (p^+) and a nega-



Y* PARTICLE PRODUCTION shown in the photograph on page 782 is depicted in greater detail. The resonance particle in this case was positive; the pion produced along with it was therefore negative. Conversely, if the particle had been negative, the accompanying pion would have been positive and the second pion (from the decay of the Y*) would have been negative. The distance traveled by the Y* (broken colored line) is of the order of 10^{-13} centimeter; thus in a bubble chamber the π^+ appears to come from the point of collision.

perform the experiment a large number of times, we will obtain a continuous spectrum of speeds of the red fragment varying from zero up to some maximum. Zero speed corresponds to the case in which all the energy in the explosive charge is carried off by the other two pieces. This will not happen often, but it can happen. More likely is a fairly equal division of energy among the three pieces. Very unlikely is a sharing of all the energy between the red piece and

the other two pieces when the latter do not separate. Thus the expected speed, or energy distribution, of the observed piece is a smooth curve [see bottom illustration on page 781].

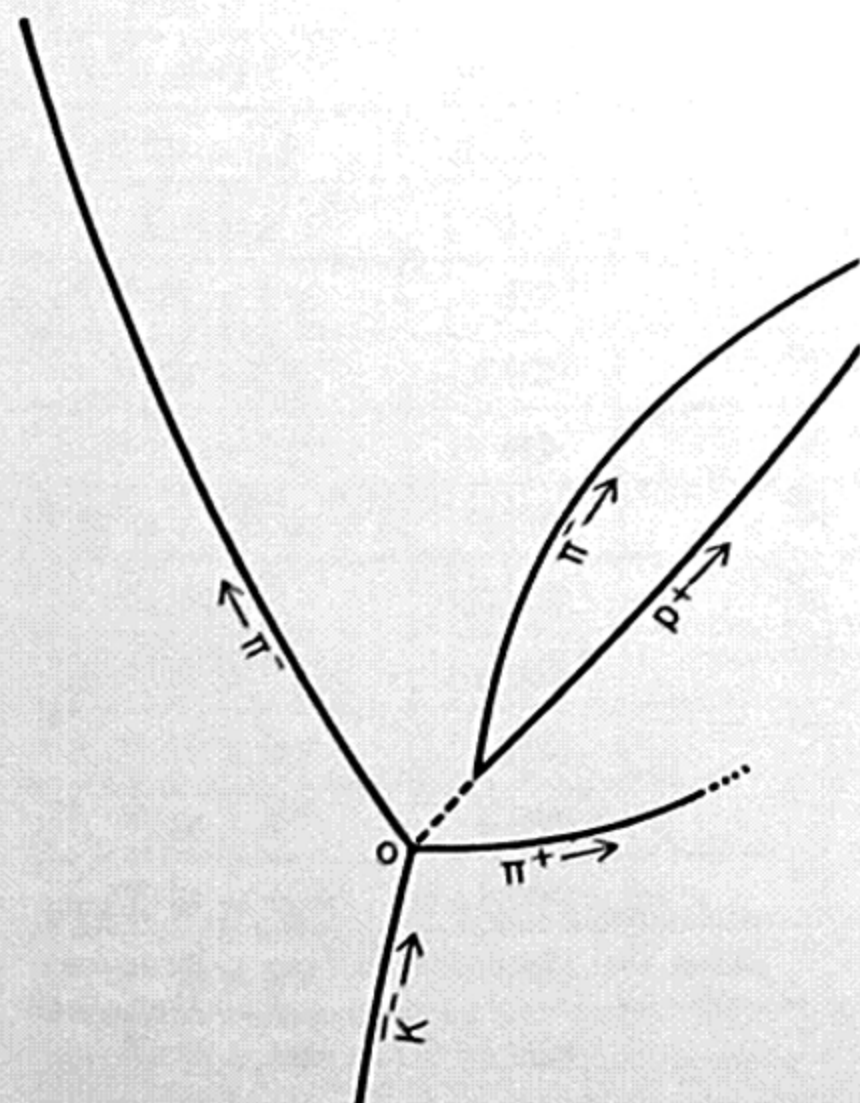
Suppose now that the two invisible pieces are cemented together with a glue that does not give way in the explosion. Then the shell breaks only into two pieces. When we observe the speed of the red piece, we find that it is always the same, because the energy of the explosion can be shared in only one way with the invisible piece. A plot of the energy is a sharp line. In this way the energy distribution of the visible piece enables us to tell whether or not the invisible section broke apart in the explosion.

The first particle experiment comparable to the one I have described was performed by the bubble-chamber group at the Lawrence Radiation Laboratory of the University of California in the summer of 1960. The Lawrence Laboratory workers were shooting a beam of high-energy negative K mesons at the liquid hydrogen in the bubble chamber. They observed that when a \bar{K}^- meson struck a proton (p), a small fraction of the collisions produced a neutral lambda particle (Λ^0) and a negative and a positive pi meson, or pion (π^- and π^+). The reaction is written: $\bar{K}^- + p^+ \rightarrow \Lambda^0 + \pi^+ + \pi^-$.

The experimental team (Margaret Alston, Luis W. Alvarez, P. Eberhard, Myron L. Good, W. Graziano, Harold

K. Ticho and Stanley G. Wojcicki) observed a few hundred of these events and with the help of a computer analyzed the energies represented by the visible pion tracks. They found a distribution with energy peaks indicating that, in a certain fraction of the selected events, one of the charged pions (plus or minus) was recoiling from one rather than two other particles [see bottom illustration on page 786]. The implication was the other pion and the lambda particle did not "break apart" immediately but remained together as a single unit at least long enough for the observed pion to recoil from it. This single unit the physicists named Y*. The reaction was envisaged as: $\bar{K}^- + p^+ \rightarrow Y^{*+} + \pi^-$. (The double plus and minus signs refer to the fact that the charge of the Y* was either + or - and was opposite to that of the pion.) In a very short time, so short that the Y* can leave no visible track, this reaction is followed by: $Y^{*+} \rightarrow \Lambda^0 + \pi^+$.

Calculations of the energies and momenta involved in the processes showed that the Y* acts like a particle with a mass of 1,384 million electron volts (Mev). This figure is made up of the rest masses of the two particles into which the Y* decays: 140 Mev for the pion and 1,115 Mev for the lambda, plus an additional kinetic energy of 129 Mev with which the pion and lambda fly apart. (According to the celebrated relation $E = mc^2$, energy and mass are equivalent quantities. In particle physics it is now customary to measure rest mass in energy units. The mass of an electron is .51 Mev; the mass of a proton, 938.2 Mev.) The reader will notice that the energy "spikes" representing the recoil from the Y* are not infinitely sharp like the spike in the projectile example. Instead they have a finite, measurable width of about 60 Mev. This width, which is made up of a spread in the energies of the various particles observed, constitutes an uncertainty in the mass-energy of the Y*. According to the uncertainty principle of quantum mechanics, uncertainty in energy is inversely proportional to uncertainty in time. An infinitely sharp spike would have an energy uncertainty of zero and therefore an infinite time uncertainty. This is the same as saying that the particle or state whose energy is represented "lives" forever; that is, it is completely stable. An energy uncertainty of 60 Mev, on the other hand, corresponds to a time uncertainty of something on the order of 10^{-23} second. This is the period



negative pion (π^-). The photograph was made by the experimental team under Luis W. Alvarez at the Lawrence Radiation Laboratory.

within which the Y^* can exist as a separate entity; in other words, it is a measure of its average lifetime.

What exactly is the Y^* ? Is it a pion and a lambda particle traveling briefly together before they take separate paths? Or is it an elementary particle that turns into a pion and a lambda particle in about 10^{-23} second? No one really knows. It may even be that with such short lifetimes the distinction is not meaningful. Whatever might be the final decision on this point, the current usage is to describe the Y^* as a resonance particle. The basis of the term "resonance" is as follows.

After the Y^* had been found its discoverers at once pointed out its similarity to a previously known resonance: the resonance between the pion and the proton (or neutron). According to the present convention the resonance could be called an N^* , N representing a nucleon (proton or neutron). This had been discovered in quite a different way. In 1952 Enrico Fermi and his colleagues at the University of Chicago were carrying out experiments in which a beam of pions was scattered by protons. They, and somewhat later other workers at the Carnegie Institute of Technology, found that the cross section, which is a measure of the probability, of scattering increased sharply beginning at a beam energy of about 100 Mev and continued to increase up to nearly 200 Mev, the highest-energy pion beam available from the accelerators of the time. When the three-billion-electron-volt Cosmotron at the Brookhaven National Laboratory went into operation, Luke C. L. Yuan and Seymour J. Lindenbaum were able to show that the cross section reached a distinct peak at a pion energy of 195 Mev and then fell off quite sharply again.

Keith A. Brueckner, then at Indiana University, suggested that there was an unusually strong and characteristic interaction between a pion and a proton that caused them to have a resonance at this energy. The characteristic feature of such an N^* resonance is the phase, or relative timing, of the oscillations of the wave associated with the scattered pion. (Thinking of particles as waves is, of course, always permissible in quantum mechanics. The probabilities of pion-scattering at different energies are related to the changes of phase of the pion waves as they pass by and through the nucleon. The amounts of phase change can be inferred from scattering observations.) As in many other resonant vibrat-

ing systems that are encountered in physics, the phase of the scattered pion wave is shifted a quarter wavelength, or 90 degrees, at resonance, and the angle measuring the amount of shift increases and decreases smoothly on both sides of the resonance point [see bottom illustration on page 788]

Here too arises the question: What is the physical interpretation of the resonance? Do the target proton and the incident pion temporarily merge into a single fundamental particle—the N^* —when they are close together at the right energy or do they retain their individuality and merely interact (for example, whirl about each other) very strongly? Again the answer is uncertain, but again the answer may be meaningless. In any case the N^* behaves like a particle with a rest mass of 1,237 Mev and a lifetime even a little shorter than that of the Y^* .

In the years from 1952 to 1960 much effort went into analyzing the nature of the pion-nucleon resonance at 195 Mev and also into a search for additional resonances in the pion-scattering cross section at higher energies. Several more resonances have in fact been found, and their characteristics are now known to be different from those of the original

resonance. To understand wherein lies the difference it is necessary to become acquainted with two rather technical concepts of particle physics.

One of these is reasonably straightforward in that it has an analogy in classical physics. It is angular momentum. Many fundamental particles have an intrinsic angular momentum, or spin. (Some have no spin.) In all cases a measurement of the amount in a preferred direction is quantized: it may in a particular case be $+1/2$ or $-1/2$ unit (the plus sign refers to spin in one direction; the minus, to spin in the opposite direction), $+1$, 0 or -1 , $+3/2$ or $-3/2$ units and so on, but never other than integral or half-integral values. Continuing the classical particle analogy still further, in a system of two particles that revolve around each other the orbital motion gives them additional angular momentum, which is also quantized. The total angular momentum of a system of two particles consists of the sum of the spin and the orbital angular momentum. Depending on their relative directions the two may add to or subtract from each other.

The pion has no intrinsic angular momentum, or spin; the nucleon has a spin of $1/2$. Analysis of the interactions of

	PARTICLE	PARTICLE CHARGE STATES	ANTIPARTICLE CHARGE STATES	MASS (MEV)	MEAN LIFE (SECONDS)
LEPTONS	NEUTRINO	ν_e, ν_μ	$\bar{\nu}_e, \bar{\nu}_\mu$	0	STABLE
	ELECTRON	e^-	e^+	.51	STABLE
	MUON	μ^-	μ^+	105.66	2.2×10^{-6}
BOSONS	PHOTON	γ	γ	0	STABLE
	PION	π^0	π^0	135	2.3×10^{-16}
		π^+	π^-	139.6	2.6×10^{-8}
	K-MESON	K^+	\bar{K}^-	494	1.2×10^{-8}
		K^0	\bar{K}^0	497.8	$6 \times 10^{-8} \quad 1 \times 10^{-10}$
BARYONS	PROTON	p^+	\bar{p}^-	938.2	STABLE
	NEUTRON	n^0	\bar{n}^0	939.5	1×10^3
	LAMBDA	Λ^0	$\bar{\Lambda}^0$	1115.4	2.5×10^{-10}
	SIGMA	Σ^+	$\bar{\Sigma}^-$	1189.4	$.8 \times 10^{-10}$
		Σ^0	$\bar{\Sigma}^0$	1191.5	$< .1 \times 10^{-10}$
		Σ^-	$\bar{\Sigma}^+$	1196	1.6×10^{-10}
	XI	Ξ^0	$\bar{\Xi}^0$	1311	1.5×10^{-10}
Ξ^-		$\bar{\Xi}^+$	1318.4	1.3×10^{-10}	

TYPE I PARTICLES are considered fundamental particles by virtue of their relatively long lifetimes, which average a ten-billionth of a second (10^{-10} second). On the basis of their masses the particles on this list have been classified as leptons, bosons or baryons.

pions and nucleons shows that their orbital angular momentum in the N^* state is 1. Depending on the relative directions, the spin of the nucleon could combine with the orbital angular momentum so as to add or subtract from it and give a total spin of $1/2$ or $3/2$. In the case of the pion-nucleon resonance at 195 Mev the two apparently add to give an angular momentum of $3/2$. If the resonance is thought of as a single particle, the two components can be considered to have merged into the spin of this particle. If the resonance is thought of as

composite, there is a mixture of spin and orbital momentum.

The second concept used to classify the resonances bears the rather deceptive name of isotopic spin. The name is misleading because the word "spin" is used largely in a figurative sense. Isotopic spin is actually a series of quantum numbers, like those describing real spin, which describe the possible charge states of a particle. For example, a particle such as the lambda, which is always neutral, is said to have an isotopic spin of zero. The nucleon, which is either the positive

proton or the neutral neutron, is assigned isotopic spin of $1/2$. A spin of $+1/2$ corresponds to positive charge; a spin of $-1/2$, to neutral charge. Some particles, including the pion, have three possible charge states: positive, negative and neutral. The isotopic spins corresponding to these are $+1$, -1 and 0 . When two particles form a resonant system, their isotopic spins are "aligned" in such a way as to add or subtract. In the case of the pion-nucleon resonance at 195 Mev the pion isotopic spin of 1 is added to the nucleon isotopic spin of $1/2$ to give a total of $3/2$. Therefore the resonance is called a $3/2, 3/2$ state; that is, the isotopic spin is $3/2$ and the angular momentum is $3/2$.

As described above isotopic spin might seem to be an arithmetical label for charge. Actually it is more than that. The quantum numbers have a deeper physical significance that can only be hinted at here. It turns out that the probabilities of various reactions that are otherwise equivalent depend sensitively on isotopic spin. Specifically the theory predicts that the proton-scattering cross section at the N^* resonance for positive pions and protons should be three times the cross section for negative pions and protons. As can be seen in the illustration on page 787, this ratio is found almost exactly in the experiments.

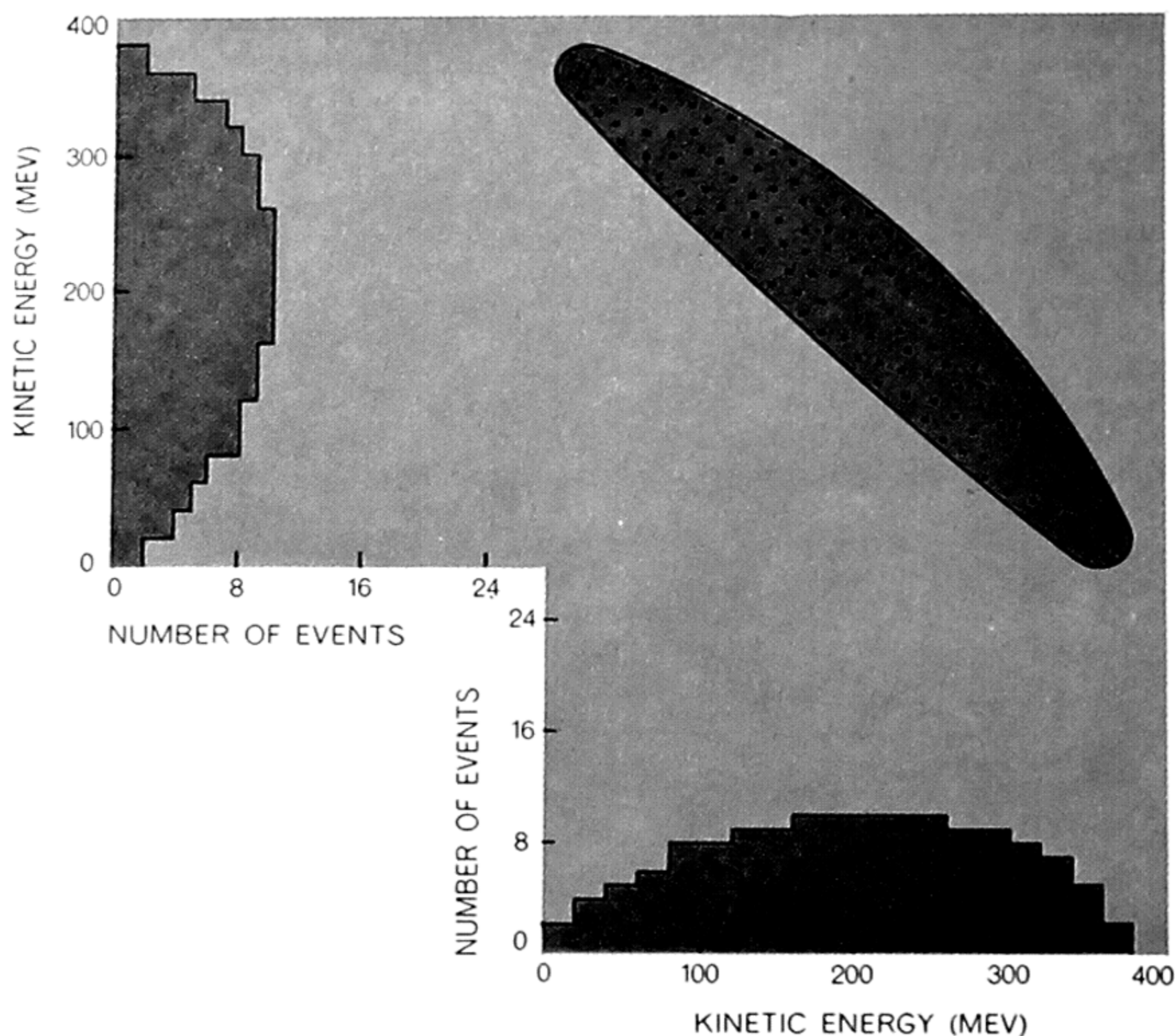
Ordinary spin and isotopic spin, then, are two of the most important properties of any particle or resonance particle. A knowledge of these quantities makes it possible to predict many of the reactions in which the particle may participate. As the illustration at the left shows, a large number of resonances have now been found. For some the ordinary spin and isotopic spin have been definitely determined. For others the values are still doubtful or, in a few cases, unknown.

The present article aims at no more than a "phenomenological" description of the resonance particles and not at a theoretical interpretation. Many theorists are busy trying to find schemes to account for them, but I shall mention these attempts only briefly.

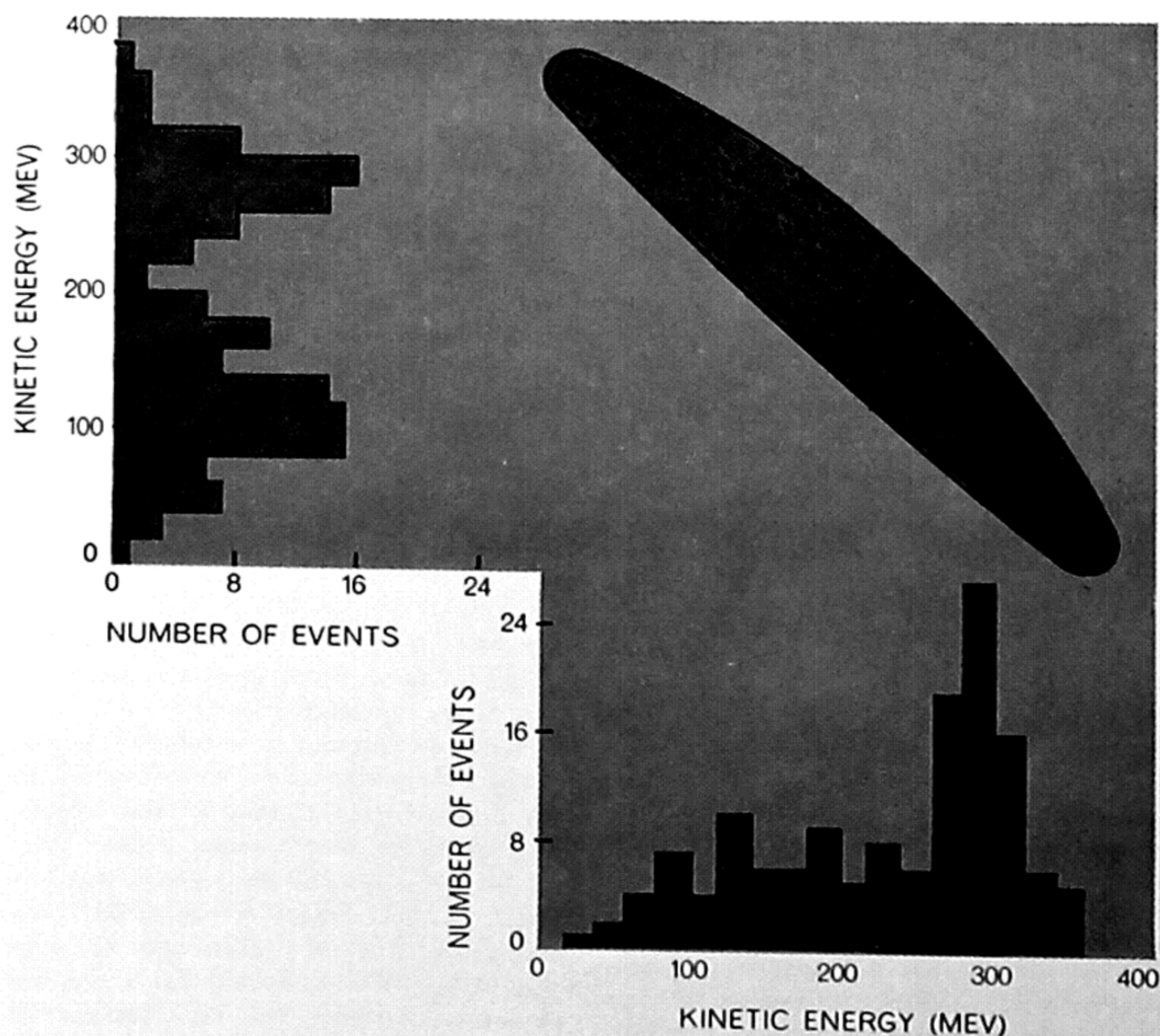
Some of the resonances were in fact predicted before they were found. In the case of the original Y^* the experimenters were led to analyze their data in the way they did by an idea put forward by Murray Gell-Mann of the California Institute of Technology. He suggested that there should be a general symmetry among the interactions of pions with all the various "baryons"—

RESONANCE PARTICLE	ISOTOPIC SPIN	TOTAL ANGULAR MOMENTUM	MASS (MEV)	PARTICLE PRODUCTION
$\eta(\pi^+\pi^-\pi^0)$	0	0	550	$\pi^+ + d^+ \rightarrow \eta^0 + p^+ + p^+$
$\rho(\pi\pi)$	1	1	760	$\rho^+ + \bar{p}^- \rightarrow \rho^0 + \pi^+ + \pi^-$ $\pi^\pm + p^+ \rightarrow \rho^\pm + p^+$
$\omega(\pi^+\pi^-\pi^0, \pi^0\gamma)$	0	1	790	$\rho^+ + \bar{p}^- \rightarrow \omega^0 + \pi^+ + \pi^-$ $\pi^+ + d^+ \rightarrow \omega^0 + p^+ + p^+$
$K^*(K\pi)$	$1/2$	1	880	$\bar{K}^- + p^+ \rightarrow K^* + p^+$ $\pi^- + p^+ \rightarrow K^* + \Sigma$
$K\bar{K}$?	?	1,020	$\pi^- + p^+ \rightarrow \bar{K}^0 + K^0 + n^0$ $\bar{K}^- + p^+ \rightarrow \bar{K}^- + K^+ + \Lambda^0$
$N^*(\pi N)$	$3/2$	$3/2$	1,237	$\pi^\pm + p^+ \rightarrow \pi^\pm + p^+$
$Y^*(\pi\Lambda, \pi\Sigma)$	1	$3/2$	1,384	$\bar{K}^- + p^+ \rightarrow Y^* + \pi$ $\pi^- + p^+ \rightarrow Y^* + \bar{K}$
$Y^{**}(2\pi\Lambda, \pi\Sigma)$	0	$1/2$	1,405	$\bar{K}^- + p^+ \rightarrow Y^{**} + \pi$
$N^{**}(\pi N)$	$1/2$	$3/2$	1,516	$\pi^- + p^+ \rightarrow N^{**} + \pi$
$Y^{***}(\pi\Lambda, \pi\Sigma, K\Lambda)$	0	$3/2$	1,520	$\bar{K}^- + p^+ \rightarrow Y^{***} + \pi$
$\Xi^*(\pi\Xi)$	$1/2$	$> 1/2$	1,535	$\bar{K}^- + p^+ \rightarrow \Xi^* + K$
$N^{***}(\pi N)$	$1/2$	$5/2$	1,683	$\pi^- + p^+ \rightarrow N^{***} + \pi$

TYPE II PARTICLES are the resonance particles. Those listed here are considered to be reasonably well established, but many of the values given are still tentative. No generally accepted nomenclature yet exists. Four are identified by Greek letters: eta (η), rho (ρ), omega (ω) and xi* (Ξ^*). The decay particles are shown in parentheses. For example, an omega particle decays into either three pions or a pion and a photon, the K^* decays into a K meson and a pion, the N^* into a pion and a nucleon (i.e., proton or neutron), and so on. The column at far right represents the reactions that produce the various particles; d represents a deuteron, or the nucleus of a heavy-hydrogen atom, consisting of one proton and one neutron. In contrast to Type I particles, the resonance particles have much shorter life-times, of the order of a hundred-thousandth of a billion-billionth of a second (10^{-23} second).



EXPECTED DALITZ PLOT shows the theoretical distribution of energy between the two pion products of the reaction: $\bar{K}^- + p^+ \rightarrow \Lambda^0 + \pi^+ + \pi^-$. The energy of the positive pion produced in 141 instances of this reaction can be read on the horizontal scale at bottom; the energy of the negative pion produced in the same events, on the vertical scale at upper left. The distribution of energy between the two pions in a large number of events should be more or less equal, and the energy plot for any one product should result in the type of histogram shown, which is the equivalent of a smooth curve (see illustration on page 781). The lenticular area at upper right defines the range of values within which the energies from a single event must fall. The distribution of energies (*dots*) is uniform.



OBSERVED DALITZ PLOT of the interaction of a negative K meson and a proton is based on the studies of the bubble-chamber group under Alvarez. The energy distribution in the lenticular area is not uniform, and the plots show that the distribution of energy occurs mainly in peaks: two relatively strong peaks in the plot of the negative pion and a single stronger peak in that of the positive pion. The distribution is consistent with a reaction that produces two particles rather than three: $\bar{K}^- + p^+ \rightarrow Y^* + \pi$. The width of the strongest resonance peak is 60 Mev, which corresponds to an average lifetime for the Y^* of 10^{-23} second. This type of graphic analysis received its name from Richard H. Dalitz of the University of Chicago, who developed it to study tau-meson decay.

particles as heavy as nucleons or heavier. Since a resonance between pion and nucleon was already known, this suggested that there should also be a resonance between the pion and the lambda particle, which is one of the baryons.

Another line of theoretical work has led to the discovery of several resonances among pions. This work got its impulse from recent studies of the scattering of electrons by nucleons. Robert Hofstadter and his colleagues at Stanford University were the pioneers in this field [see "The Atomic Nucleus," by Robert Hofstadter; *SCIENTIFIC AMERICAN* Offprint 217], and they were later joined by Robert R. Wilson's group at Cornell University and G. R. Bishop's group at the Orsay laboratory of the French National Center for Scientific Research. Their experiments have shown that both the electric and the magnetic properties of protons and neutrons are not concentrated at a point but are distributed over a space of finite size. In other words, the experiments are depicting the electromagnetic structure of the nucleon.

That structure turns out to be analyzable into three separate parts. First, there is a core: a small, central region of

positive charge that accounts for about a fourth of the total charge. Second, there is a "vector" portion that is positive in the proton and negative in the neutron and that extends over the whole nucleon; it accounts for about half of the total charge. Third, there is a positive "scalar" portion, also extending over the whole particle and contributing a fourth of the total charge.

As for magnetism, part of it in the case of the proton is directly identified with the spinning charge. But there is another part, which occurs in both proton and neutron, that cannot be identified with the net over-all charge. This is known as the anomalous magnetic moment. It too turns out to consist of three components resembling those of the charge.

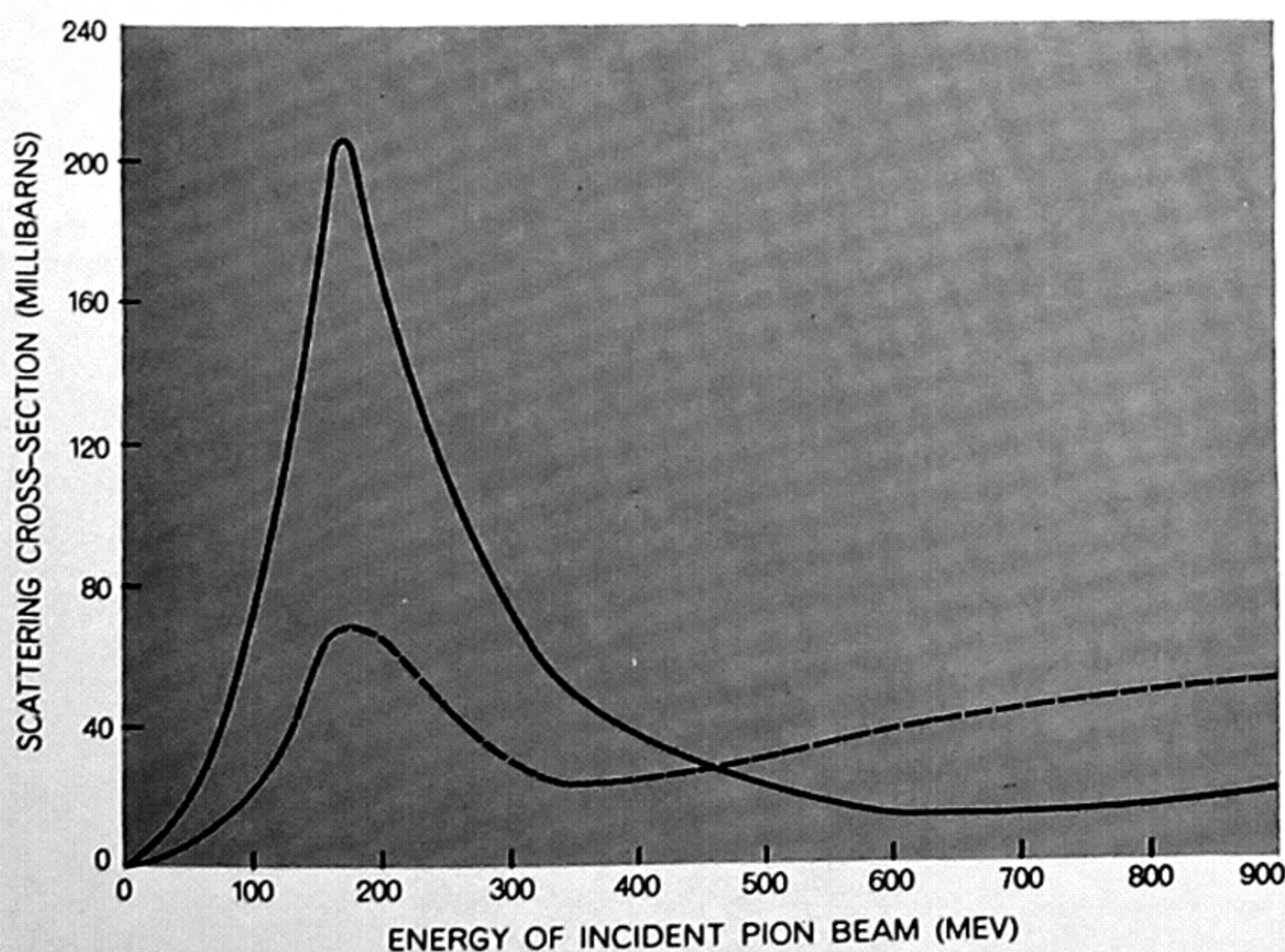
When the detailed pictures of the electromagnetic structure of the nucleon began to emerge, it was at once apparent that they would not fit satisfactorily into the then prevailing theory of the nucleon. All such theories are based on still another concept peculiar to quantum theory: the idea of virtual "field particle" emission. Briefly, it is believed

that a proton or neutron continually emits and reabsorbs virtual pions. The time that the nucleon spends in this virtual state and the distance that the pion separates from the nucleon are consistent with the uncertainty principle. Thus for a certain fraction of the time the core of the nucleon is surrounded by a meson cloud, and it is this cloud that accounts for the extended charge and magnetic moment.

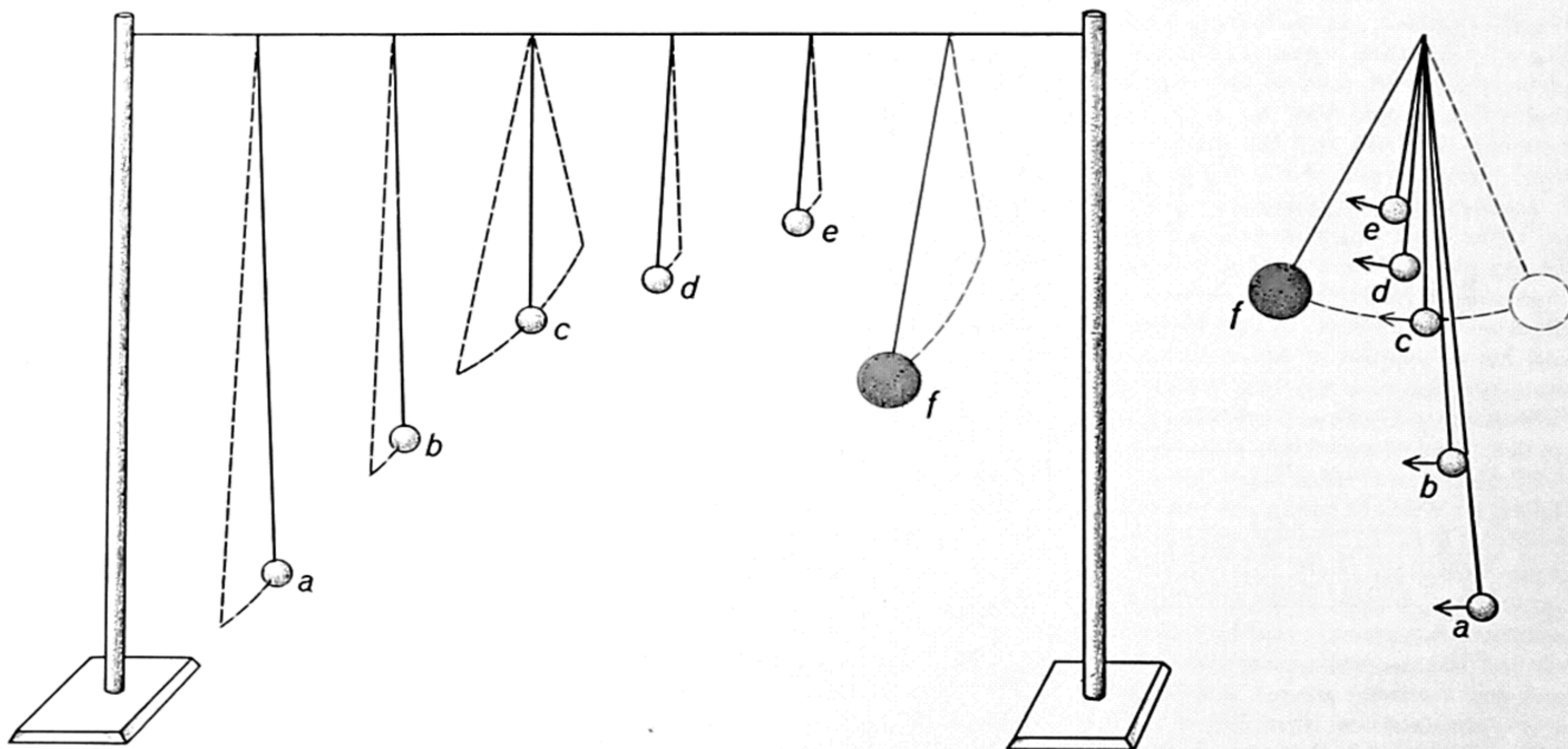
Originally the virtual-emission process was envisaged in terms of noninteracting pions. But Hofstadter's results could not be explained in this way. In 1959 William R. Frazer and Jose R. Fulco of the Lawrence Radiation Laboratory showed that the vector part of the charge and the magnetic properties would be accounted for if the nucleon emitted two pions and these entered into a strong, or resonant, interaction while they were out in the cloud. To put it another way, the calculations predicted a two-pion resonance particle, and they showed that it should have an isotopic spin of 1, an angular momentum of 1 and a mass of approximately 600 Mev.

These calculations prompted experimenters to look for a 600-Mev resonance particle in various pion-pion interactions. Evidence for it was soon found independently by a number of groups. One result, for example, showed that when a high-energy pion produces a second pion by colliding with a proton, there is a strong attraction between the two pions. Eventually experiments analogous to those outlined for the Y^* demonstrated that the resonance known as the rho particle has a mass of approximately 760 Mev, which is in good agreement with the current theory of electromagnetic structure of the nucleon.

Also in 1959 Geoffrey F. Chew of the Lawrence Radiation Laboratory pointed out that the scalar part of the nucleon electromagnetic structure could be understood in terms of another resonant interaction, this time involving three pions. Since this part of the nucleon structure is the same for proton and neutron, the isotopic spin of a three-pion resonance interaction needed to explain this feature has to be zero; that is, it must exist in only one neutral charge form. Chew suggested that it was not unreasonable to anticipate the existence of a strong three-pion resonance particle, which should have a mass approximately the same as the two-pion resonance state. It should be mentioned that Chew's suggestion of a zero isotopic-spin particle was not the first. Two years earlier



PION-PROTON RESONANCE, discovered in 1952, was the first of its kind. The two curves plot the probability that a beam of positive (*solid line*) and negative pions (*broken line*) will be scattered by protons. The scattering cross section, or probability (measured in millibarns), for positive pions begins to increase sharply at about 100 million electron volts (Mev), reaches a peak of somewhat more than 200 Mev and then falls off almost as sharply. In contrast, the resonant effect for the scattering of negative pions is only a third as strong.



COUPLED PENDULUMS provide a mechanical illustration of a resonant system. The frequency of the driver pendulum (*f*) is greater than that of the first two "slave" pendulums (*a* and *b*), the same as that of the third (*c*) and smaller than that of the last two

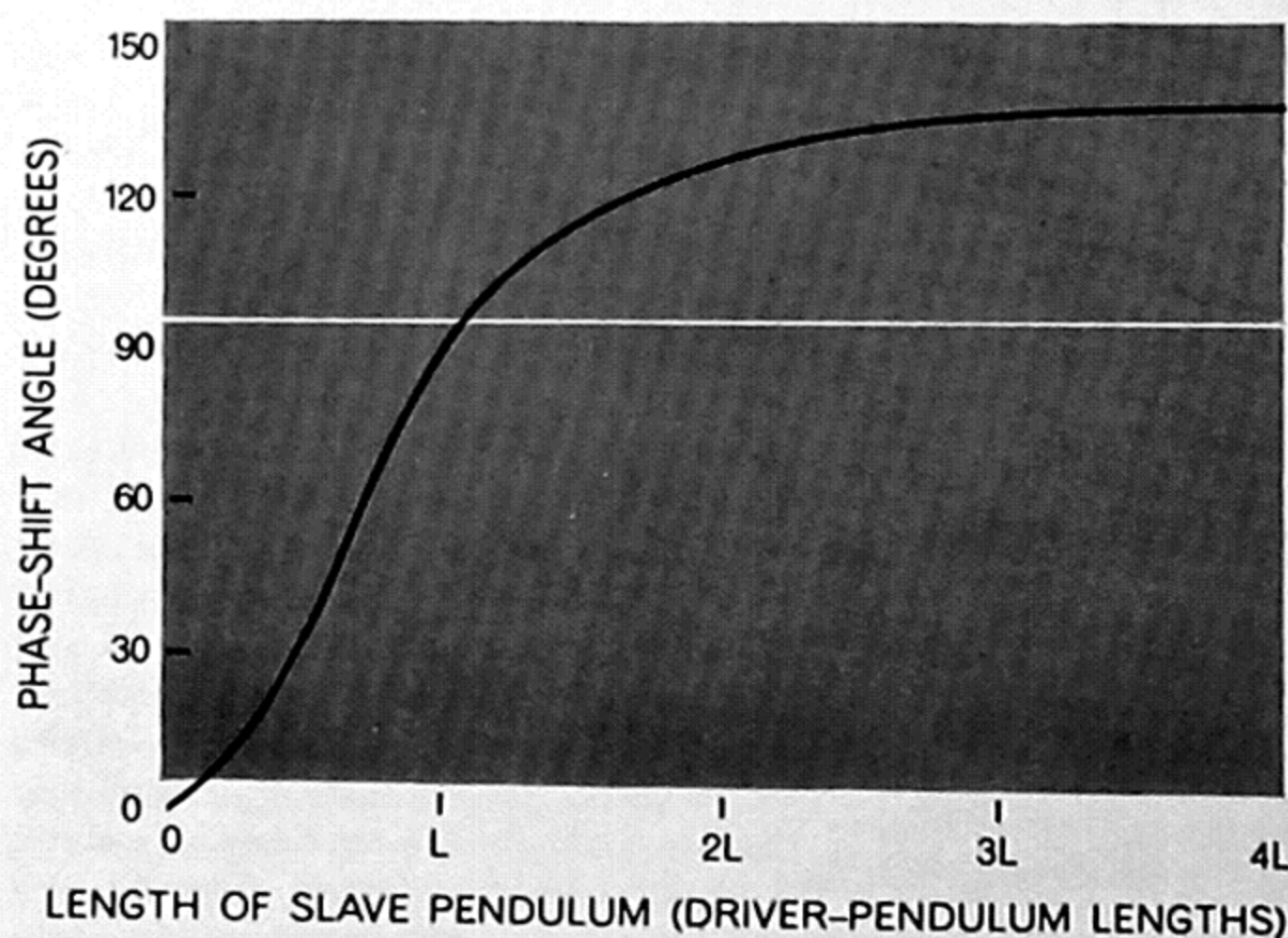
(*d* and *e*). The third oscillates 90 degrees out of phase with the driver (i.e., a quarter of a cycle behind it). The first two oscillate between 90 and 180 degrees behind, the last two between zero and 90 degrees behind. This lag is the phase shift (see illustration below).

Yoichiro Nambu of the University of Chicago had suggested the existence of a neutral heavy meson that would contribute to the electromagnetic structure of the nucleon. These ideas were clearly responsible for the research that led to the discovery of the various multiple-pion resonance states now known. Probably the most spectacular experimental discovery was that of the omega: the three-pion resonance that Chew and

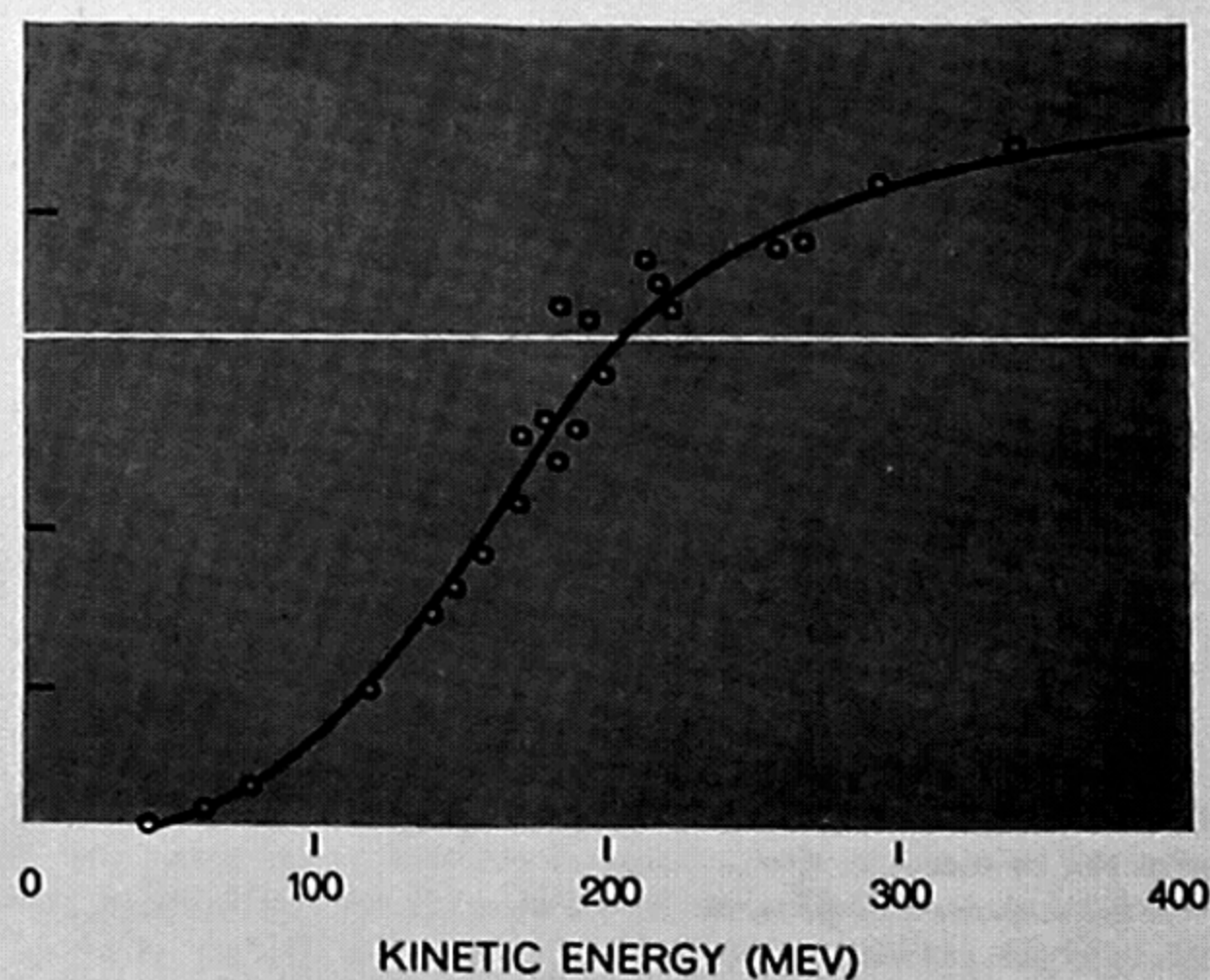
Nambu had predicted. The experiments were carried out by B. C. Maglic, Luis W. Alvarez, Arthur H. Rosenfeld and M. Lynn Stevenson of the Lawrence Laboratory. They studied the annihilation of antiprotons (\bar{p}) encountering protons in the 72-inch liquid-hydrogen bubble chamber. Annihilations yield a wide variety of products, but the experimenters concentrated only on those that produced four outgoing pion tracks. Out

of 2,500 events of this type only 800 had the following special property: in order to balance energy and momentum in the annihilation process, another neutral pion must have been present with the outgoing particles. These carefully restricted events were therefore examples of the following reaction: $p^+ + \bar{p}^- \rightarrow \pi^+ + \pi^- + \pi^0 + \pi^+ + \pi^-$.

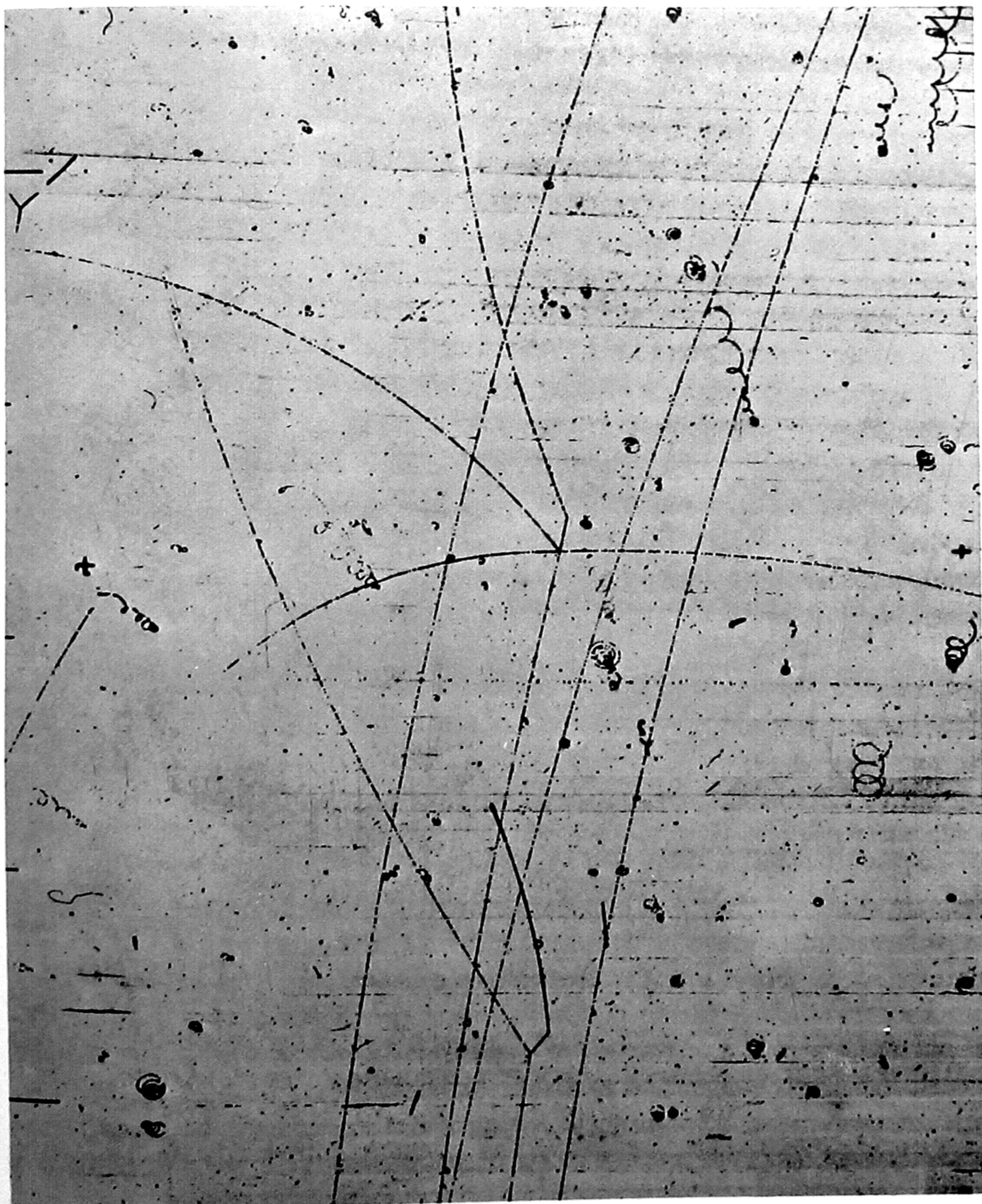
In all 800 examples of the reaction the physicists hoped to find some energy



RESONANCE EFFECT in the scattering of pions by protons becomes apparent when energy is plotted against phase shift (graph

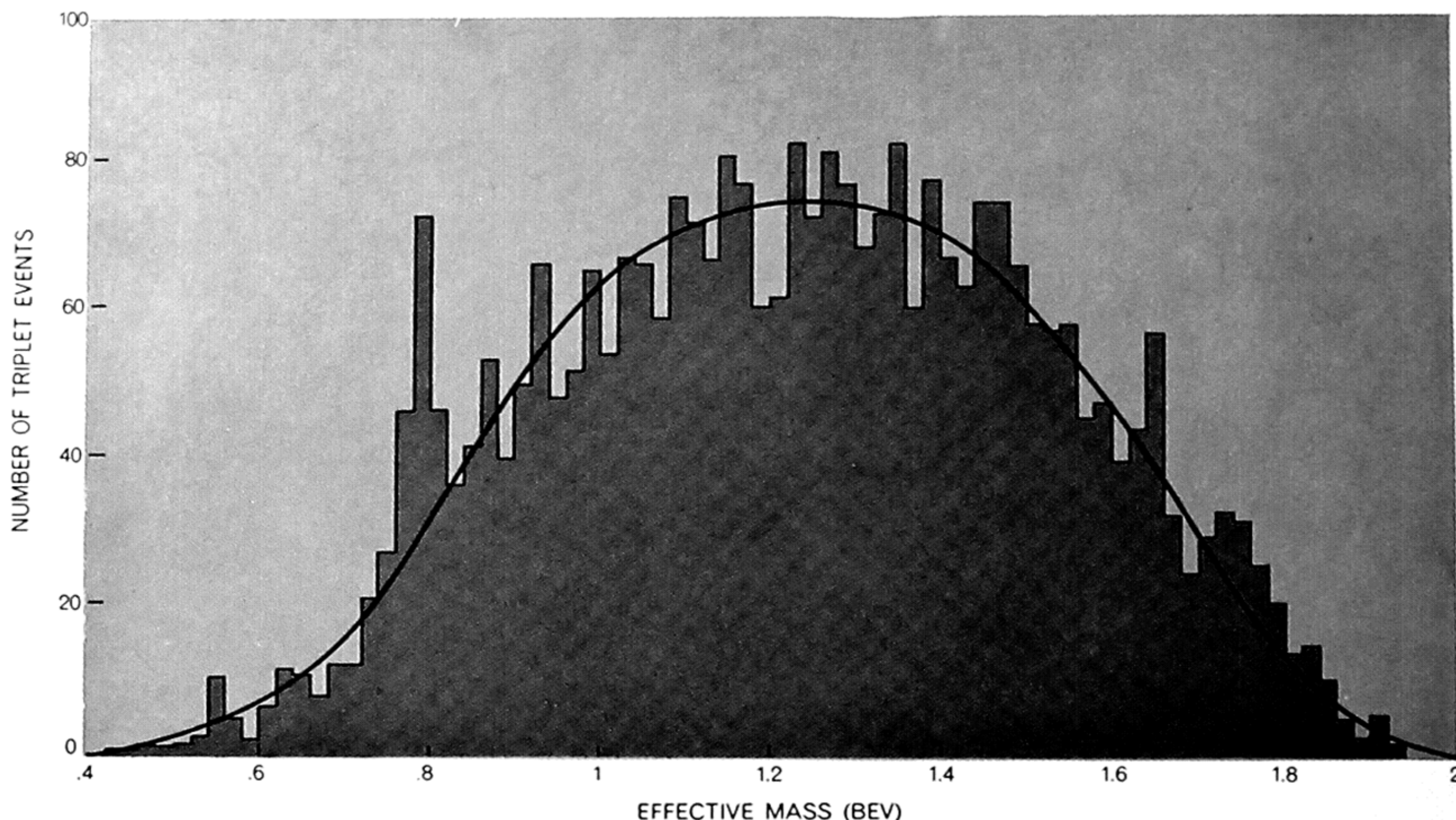


at right). The curve closely resembles that of the corresponding plot of phase angle against the pendulum length (graph at left).



PRODUCTION OF POSSIBLE Y^{**} from the collision of a negative K meson (\bar{K}^- in drawing at right) with a proton at O cannot be distinguished from the direct production of a negative sigma

particle (Σ^-), one negative pion and two positive pions (and possibly even a neutral pion) in this reaction, unless careful analyses are made of the energies and momenta of the particles involved.



OMEGA RESONANCE PARTICLE, a three-pion particle, was discovered in experiments carried out at the Lawrence Radiation

Laboratory. Its observed mass (*peak near .8 Bev*) was 790 Mev. All the other masses tended to average out along the smooth curve.

where three of the pions exhibited a resonant interaction. (Since they were looking for a neutral resonance, they knew that one of the pions would have to be neutral and the other two oppositely charged.) A further detailed analysis of the dynamics of the 800 events was

then performed (using high-speed digital computers throughout) and an "effective" mass was computed for every possible combination of three of the five pions present in each event. The method was similar to that already discussed in connection with the discovery of the Y^0 .

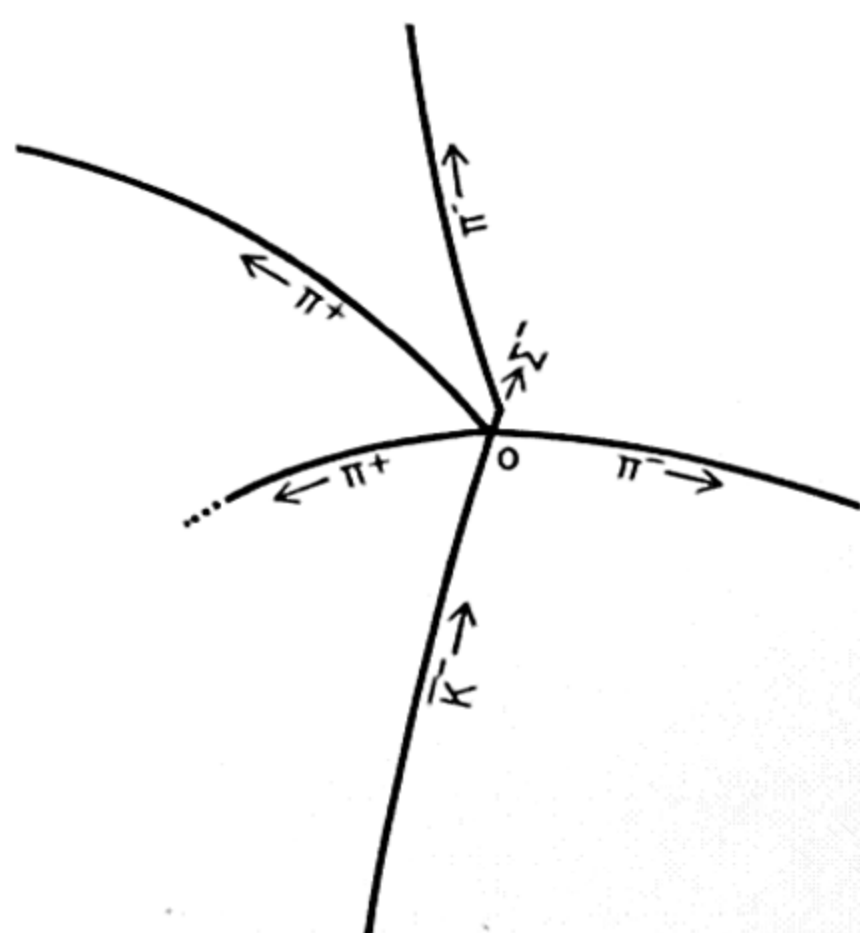
It was found that only the expected combination—namely, the π^+ , π^- , π^0 combination—led to a group of mass values having a peak value characteristic of a single particle [see illustration above]. Very few of the pion triplets, in fact only 93, showed the property of being in the resonant state. Presumably almost all the other cases represented production of independent pions. Nevertheless, there were enough to point clearly at the omega particle (ω^0), and to show that it is a neutral combination (π^+ , π^- , π^0) arising from the reaction: $p^+ + \bar{p}^- \rightarrow \omega^0 + \pi^+ + \pi^-$. The observed mass of the omega resonance particle is 790 Mev, and its lifetime, as determined from the width of the particle resonance, is equal to or greater than 4×10^{-23} second.

At present the nature of the resonance particles is very much in question. As has been mentioned, certain theoretical ideas account fairly well for certain of

the particles. But there are some particles that seem to have no such pedigree.

Of course, theoretical physicists are trying hard to find some general framework that will accommodate all the resonances. There have been several independent lines of attack, which I can do no more than to identify in a few words. One has been to regard some of the resonance particles as the quanta of certain fields, just as the photon is the quantum of the electromagnetic field and the pion the quantum of the nuclear-force field. Another approach has been to consider that all possible particles are associated with a representation of the mathematical form known as a group. A third line makes use of a new idea in the mathematics of quantum theory known as Regge poles. Here all particles are regarded as equally fundamental and equally composite, each being representable as a dynamical interaction of the others.

On this necessarily mysterious note the article closes. If the reader is mystified, so are physicists. The place of the resonance particles in the scheme of things is one of the most puzzling physical questions to which the future, one hopes, will provide the answer.



The sigma particle decays into a negative pion and a neutron, which leaves no track. Photograph was made by the Alvarez group.

The Author

R. D. HILL is professor of physics at the University of Illinois, where since 1947 he has taught and done experimental research. For the past eight years his research has been mainly in the field of high-energy nuclear physics. He was associated with the group at the Brookhaven National Laboratory that first detected in nuclear emulsions the production of tau and K mesons by the Brookhaven Cosmotron in 1954. Before going to the University of Illinois, Hill, who was born and raised in Australia, had been senior lecturer at the University of Melbourne since 1941. In the years 1939 to 1945 he was involved in radar work, first in England and later in Australia. Prior to that Hill took his D.Sc. at Melbourne in 1936, did research for a year at the University of Cambridge and worked the following year at Illinois.

Bibliography

EVIDENCE FOR A $T=0$ THREE-PION RESONANCE. B. C. Maglić, L. W. Alvarez, A. H. Rosenfeld and M. L. Stevenson in *Physical Review Letters*, Vol. 7, No. 5, pages 178-182; September, 1961.

MESONS AND HYPERONS. C. A. Snow and M. M. Shapiro in *Reviews of Modern Physics*, Vol. 33, No. 2, pages 231-238; April, 1961.

PION-HYPERON RESONANCES. Margaret H. Alston and Massimiliano Ferro-Luzzi in *Reviews of Modern Physics*, Vol. 33, No. 3, pages 416-426; July, 1961.

SPECIAL NOTE TO TEACHERS: Each article in this volume, plus more than 660 others, is available as a separate, self-bound SCIENTIFIC AMERICAN Offprint. Offprints may be ordered in any combination and in any quantity. Teachers who want to adopt articles for their courses, therefore, can ensure that each student has his own set. Students' sets are collated by the publisher before shipment.

DATE LOANED

Acc. No. _____

[illegible]

DATE LOANED

Book No. _____

Acc. No. _____

[illegible]

DATE LOANED

Book No. _____

Acc. No. _____

[illegible]

DATE LOANED

Book No. _____

Acc. No. _____

This book may be kept for **14 days**. An over - due charge will be levied at the rate of **10 Paise** for each day the book is kept over - time.

[illegible]